Lossless coding for distributed streaming sources

Cheng Chang, Stark C. Draper, and Anant Sahai
cchang@eecs.berkeley.edu, sdraper@eecs.berkeley.edu, sahai@eecs.berkeley.edu

Abstract

Distributed source coding is traditionally viewed in the block coding context — all the source symbols are known in advance at the encoders. This paper instead considers a streaming setting in which iid source symbol pairs are revealed to the separate encoders in real time and need to be reconstructed at the decoder with some tolerable end-to-end delay using finite rate noiseless channels. A sequential random binning argument is used to derive a lower bound on the error exponent with delay and show that both ML decoding and universal decoding achieve the same positive error exponents inside the traditional Slepian-Wolf rate region. The error events are different from the block-coding error events and give rise to slightly different exponents. Because the sequential random binning scheme is also universal over delays, the resulting code eventually reconstructs every source symbol correctly with probability 1.

This material was presented in part at the IEEE Int Symp Inform Theory, Adelaide, Australia, Sept 2005.

Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, CA 94720

Mitsubishi Electric Research Labs in Cambridge, MA. This work was performed while he was a postdoc at Wireless Foundations in the University of California Berkeley.

Wireless Foundations, Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, CA 94720

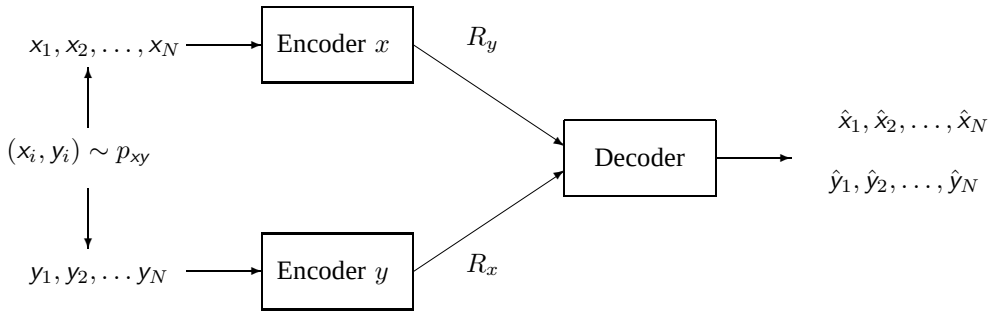

Fig. 1. Slepian-Wolf distributed encoding and joint decoding of a pair of correlated sources.

Lossless coding for distributed streaming sources[1]

I. Introduction

Traditionally, "lossless" coding is considered using two distinct paradigms: fixed block coding and variable-length coding[2]. As classically understood, both consider that the source-symbols are known in advance at the encoder and that they must be mapped into a string of bits decoded by the receiver. Fixed-block coding accepts a small probability of error and constrains the length of the bit-string, while variable-length encoding constrains only the *expected* length of the bit-string in exchange for keeping the probability of error at zero. In the point-to-point setting, both paradigms apply generically. In contrast, distributed source coding, has traditionally been explored within the fixed block context. In [1], Slepian and Wolf even asked:

> What is the theory of variable-length encodings for correlated sources?

In the classical context of source realizations known entirely in advance, the answer is simple: there is no nontrivial sense of variable-length encoding that applies generically while still being interesting.[3] This is easiest to see by example (Illustrated in Figure 1 and revisited as Example 2 in Section IV). Suppose that the first encoder observes the random vector $\mathbf{x}$, which consists of a sequence of N iid uniform binary random variables. Suppose further that the second encoder observes $\mathbf{y}$ which is related to $\mathbf{x}$ via a memoryless binary symmetric channel with crossover probability $\rho < 0.5$. The Slepian-Wolf sum-rate bound is $H(x, y) = 1 + H(\rho) < 2 = H(x) + H(y)$. But since the individual encoders only see uniformly distributed binary sources, they do not know when the sources are behaving jointly atypically. Therefore, they have no basis on which to adjust their encoding rates to combat joint atypicality. Since all pairs are possible when finite blocklengths are considered, the individual encoders must use distinct bit-strings for each of them. Since the expected length depends only on the uniform marginal distributions, this means that the expected length must be at least N. Thus, variable-length approaches do not, in general[4], lead to zero-error Slepian-Wolf codes for interesting rate-points.

Another view of variable-length coding is as a tool that enables us to achieve meaningful compression despite not knowing the underlying probability distribution[5] and allowing the rate used to adapt to the source. If there is

[1]This material was presented in part at the IEEE Int Symp Inform Theory, Adelaide, Australia, Sept 2005.

[2]There are actually four different traditional cases: fixed to fixed, fixed to variable, variable to fixed, and variable to variable. However, the last three all achieve a probability of error of zero and so we consider them together.

[3]At least at sum rates close to the joint source entropy rate. If the rates of communication are high enough, e.g., equaling the log of the cardinalities of the source alphabets, zero-error communication is possible.

[4]One should note that, in analogy to zero-error channel coding, there are special (non-generic) cases where zero-error Slepian-Wolf coding is possible [2] since certain symbol pairs cannot occur.

[5]In the point-to-point case, this is very closely related to achieving a zero-error probability. The same string can be an atypical realization of one source model while being a typical realization of another source. Encoding all the typical sequences correctly without knowing the underlying model requires getting all the possible sequences correctly for any specific model.

a low-rate, but reliable[6], feedback link available from the decoder to the two separate encoders, then this sense of variable-length Slepian-Wolf coding is possible. [5] gives a fixed-to-variable scheme in which the stopping-time is chosen at the decoder and communicated back to the encoders over a low-rate feedback link. The goal of [5] is not achieving a truly zero probability of error — rather it is willing to accept a very small probability of error in exchange for using a rate that is as small as possible.

To answer the question posed by Slepian and Wolf in the more classical sense, we instead want to aim for a probability of error that goes to zero for every source symbol, but at the cost of a variable delay. To do this, we propose stepping back and eliminating the modeling assumption of encoders having access to the entire source realization in advance. We argue that a "streaming setting" is required to discern the system-level analog to variable-length source coding in the distributed context. The streaming setting abstracts sources that are embedded in time as well as the fact that all physically realizable encoders/decoders must obey some form of causality. Thus "rate" is not just measured in bits per source symbol but in both source symbols per second and bits per second. The source-rate (symbols per second) is specified as a part of the problem while the bit-rate (bits per second) is something that we get to choose. From an engineering perspective, three desirable qualities[7] are:

- Using a low rate bit-pipe(s)
- Low end-to-end latency
- Low probability of error

The theory of source-coding should tell us the tradeoffs between these three desiderata. In addition, we will be interested in to what extent a streaming code can be made "universal" over a class of probability distributions.

In the point-to-point streaming setting, regardless of whether block or variable-length compression is used, the traditional initial step is the same: group symbols into source blocks. To compress the data blocks, either use a fixed-rate block code, or a variable-length code. The resulting encoding is then enqueued for transmission across the bit-pipe. As long as the source entropy rate is below the data-rate, the queue will remain stable. When block coding is used for compression, there is a constant delay through the system, and atypical source blocks are received in error. The probability of error is fixed at the system's design-time and so is the end-to-end delay.

In contrast, variable-length coding induces a variable system delay. The more unlikely the source blocks, the longer the delay experienced at run-time. Thus, while *asymptotically* there are no errors when variable-length source codes are used (assuming an infinite buffer size), the delay till a given symbol can be decoded depends on the random source realization. Because atypical source realizations are large deviation events, the probability that some source symbol cannot be reconstructed Δ samples after it enters the encoder decays exponentially[8] in Δ. The choice of acceptable end-to-end delay is left to the receiver/application.

We show that this type of reliability *can* be achieved in a generic distributed coding context — the probability of error goes to zero with end-to-end delay and the choice of the acceptable delay is entirely up to the decoder. Essentially, every source symbol is recovered correctly eventually with probability[9] 1. The only difference is that unlike the point-to-point case, the decoder does not necessarily know when the estimate for the symbol has converged to its final value. Furthermore, just as in the point-to-point setting[10], both the encoding and decoding can be made universal.

In this paper, we formally define a streaming Slepian-Wolf code, and develop coding strategies both for situations when source statistics are known and when they are not. The new tool is a sequential binning argument that parallels the tree-coding arguments used to study convolutional codes. We characterize the performance of the streaming schemes through an error exponent analysis and demonstrate that the exponents are equal regardless of whether the system is informed of the source statistics (in which case we use maximum likelihood decoding) or not (in which case we use universal decoding). The universal decoder we design for the streaming problem is somewhat different from those familiar from the block coding literature, as are the nature of the error exponents.

[6]It is clear that our techniques from [3], [4] can also be adapted to make the system of [5] work using only noisy feedback channels.

[7]Of course, "implementation complexity" forms a fourth and very important consideration, but we will be ignoring that aspect of the problem.

[8]In [6], we show that variable length codes used in this manner actually achieve the best possible error exponent with delay. This is also related to the analysis of [7].

[9]The secret here is that we are considering a probability measure over infinite sequences. While all pairs of finite strings may be possible, most pairs of infinite strings collectively have probability zero.

[10]Sliding-window Lempel-Ziv compression is one example where data is naturally encoded sequentially. It is also universal over sources.

A. Potential applications and practical motivation

In addition to our core interest in answering some basic questions about Slepian-Wolf coding, our formulation is also motivated by the diverse emerging application areas for distributed source coding. Media (e.g. video-conference) sources naturally have a streaming character. Consequently, we are motivated to explore what sort of streaming Slepian-Wolf technique matches naturally to such situations.[11]

B. Outline

Section II summarizes the notation used in the paper. Section III reviews the classical block-coding error exponent results for Slepian-Wolf source coding and then we state the main results of this paper: sequential error exponents for Slepian-Wolf source coding. Section IV presents a numeric study of two example sources. We observe that the sequential error exponent is often the same as the block coding error exponent. Sections V, VI and VII prove the theorems in Section III. We start with sequential source coding for single sources in V. This is the simplest case but it provides insights to the nature of sequential source coding problem and sequential error events. We show that the sequential error exponent is the same as the random block source coding error exponent. Section VI moves on to the case with decoder side-information. Finally, Section VII presents the proof of the main result of the paper. We derive the sequential error exponent of distributed source coding for correlated sources. This error exponent strictly positive everywhere inside the achievable rate region of [1]. For all these three scenarios in Sections V, VI and VII, both ML and universal decoding rules are studied. The appendix shows that the resulting error exponents are indeed the same.

II. Notation

We use serifed-fonts, e.g., x to indicate sample values, and sans-serif, e.g., x, to indicate random variables. Bolded fonts are reserved to indicate sample or random vectors, e.g., $\mathbf{x} = x^n$ and $\mathbf{x} = \mathsf{x}^n$, respectively, where the vector length (n here) is understood from the context. Subsequences, e.g., $x_l, x_{l+1}, \ldots, x_n$ are denoted as x_l^n where $x_i^j \triangleq \emptyset$ if $i < j$. Distributions are indicated with lower-case p, e.g., x is distributed according to $p_{\mathsf{x}}(x)$. Sets and their elements are denoted as, e.g., $x \in \mathcal{X}$, and their cardinality by $|\mathcal{X}|$. We use calligraphic font to denote sets, $\mathcal{X}$, $\mathcal{F}$, $\mathcal{W}$ etc, and reserve $\mathcal{E}$ and $\mathcal{D}$ to denote encoding and decoding functions, respectively. We use standard notation for types, see, e.g., [8]. Let $N(a;\mathbf{x})$ denote the number of symbols in the length-n vector $\mathbf{x}$ that take on value a. Then, $\mathbf{x}$ is of type P if $P(a) = N(a;\mathbf{x})/n$. The type-class, or set of length-n vectors of type P is denoted $\mathcal{T}_P$. A sequence $\mathbf{y}$ has conditional type V given $\mathbf{x}$ if $N(a,b;\mathbf{x},\mathbf{y}) = N(a;\mathbf{x})V(b|a) = P(a)V(b|a)$ for every a, b. The set of sequences $\mathbf{y}$ having conditional type V with respect to $\mathbf{x}$ is called the V-shell of $\mathbf{x}$ and is denoted by $\mathcal{T}_V(\mathbf{x})$. When considered together, the pair $(\mathbf{x},\mathbf{y})$ is said to have joint type $V \times P$. We always use upper-case, e.g., P and V, to denote length-n types and conditional types. As we often discuss the types of subsequences we add a superscript notation to remind the reader of the length of the subsequence in question. If, for instance, the subsequence under consideration is x_l^n we write $x_l^n \in \mathcal{T}_{P^{n-l}}$. Similarly we use V^{n-l} for the conditional type of length-$(n-l+1)$, and $V^{n-l} \times P^{n-l}$ for the joint type.

Given a joint type $V \times P$, entropies and conditional entropies are denoted as $H(P)$ and $H(V|P)$, respectively. The KL divergence between two distributions q and p is denoted by $D(q\|p)$.

III. Main Results

In this section, we begin by reviewing classical results on the error exponents of distributed block coding. We then present the main results of the paper: error exponents for streaming Slepian-Wolf coding and its special cases: point-to-point coding and source coding with decoder side information. We analyze both maximum likelihood and universal decoding and show that the achieved exponents are equal. Leaving numerical examples and proofs for later sections, we here compare the form of the streaming exponents with their block coding counterparts.

[11] A secondary aspect in some multimedia settings is a natural multi-scale nature to the source — the high order bits are more important than the low order bits. To the extent that the high order bits can be made "early" and the low-order bits can be made "late", our constructions also naturally give more protection to the early bits as compared to the later ones. While this interpretation might eventually be important in practice, it is a bit questionable within the simplified model this paper considers.

A. Block source coding and error exponents

In the classic block-coding Slepian-Wolf paradigm, full length-N vectors $\mathbf{x}$ and $\mathbf{y}$ are observed by their respective encoders before communication commences. In this situation a rate-(R_x, R_y) length-N block source code consists of an encoder-decoder triplet $(\mathcal{E}_N^x, \mathcal{E}_N^y, \mathcal{D}_N)$, as we will define shortly. For the rate-region considerations, the general case of distributed encoders can be considered by using time-sharing among codes that alternate between sending at rates close to the marginal entropy and those that correspond to perfectly known side-information. However, it is easy to see that this results in a substantial loss of error-exponent even in the block-coding case. To get good exponents, something else is required:

Definition 1: A randomized length-N rate-(R_x, R_y) block encoder-decoder triplet $(\mathcal{E}_N^x, \mathcal{E}_N^y, \mathcal{D}_N)$ is a set of maps

$$\begin{array}{lcll} \mathcal{E}_N^x & : & \mathcal{X}^N \to \{0,1\}^{NR_x}, & \text{e.g.,} \quad \mathcal{E}_N^x(x^N) = a^{NR_x} \\ \mathcal{E}_N^y & : & \mathcal{Y}^N \to \{0,1\}^{NR_y}, & \text{e.g.,} \quad \mathcal{E}_N^y(y^N) = b^{NR_y} \\ \mathcal{D}_N & : & \{0,1\}^{NR_x} \times \{0,1\}^{NR_y} \to \mathcal{X}^n \times \mathcal{Y}^n, & \text{e.g.,} \quad \mathcal{D}_N(a^{NR_x}, b^{NR_y}) = (\hat{x}^N, \hat{y}^N) \end{array}$$

where common randomness, shared between the encoders and the decoder is assumed. This allows us to randomize the mappings independently of the source sequences.

The error probability typically considered in Slepian-Wolf coding is the joint error probability, $\Pr[(x^N, y^N) \neq (\hat{x}^N, \hat{y}^N)] = \Pr[(x^N, y^N) \neq \mathcal{D}_N(\mathcal{E}_N^x(x^N), \mathcal{E}_N^y(y^N))]$. This probability is taken over the random source vectors as well as the randomized mappings. An error exponent E is said to be achievable if there exists a family of rate-(R_x, R_y) encoders and decoders $\{(\mathcal{E}_N^x, \mathcal{E}_N^y, \mathcal{D}_N)\}$, indexed by N, such that

$$\lim_{N\to\infty} -\frac{1}{N} \log \Pr[(x^N, y^N) \neq (\hat{x}^N, \hat{y}^N)] \geq E. \tag{1}$$

In this paper, we study random source vectors $(\mathbf{x}, \mathbf{y})$ that are iid across time but may have dependencies at any given time:

$$p_{x,y}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{N} p_{x,y}(x_i, y_i).$$

For such iid sources, upper and lower bounds on the achievable error exponents are derived in [9], [8]. These results are summarized by the following theorem.

Theorem 1: (Lower bound) Given a rate pair (R_x, R_y) such that $R_x > H(x|y)$, $R_y > H(y|x)$, $R_x + R_y > H(x, y)$. Then, for all

$$E < \min_{\bar{x},\bar{y}} D(p_{\bar{x},\bar{y}} \| p_{xy}) + \left| \min[R_x + R_y - H(\bar{x}, \bar{y}), R_x - H(\bar{x}|\bar{y}), R_y - H(\bar{y}|\bar{x})] \right|^+ \tag{2}$$

there exists a family of randomized encoder-decoder mappings as defined in Definition 1 such that (1) is satisfied. In (2) the function $|z|^+ = z$ if $z \geq 0$ and $|z|^+ = 0$ if $z < 0$.

(Upper bound) Given a rate pair (R_x, R_y) such that $R_x > H(x|y)$, $R_y > H(y|x)$, $R_x + R_y > H(x, y)$. Then, for all

$$E > \min \left\{ \min_{\bar{x},\bar{y}: R_x < H(\bar{x}|\bar{y})} D(p_{\bar{x},\bar{y}} \| p_{xy}), \min_{\bar{x},\bar{y}: R_y < H(\bar{y}|\bar{x})} D(p_{\bar{x},\bar{y}} \| p_{xy}), \min_{\bar{x},\bar{y}: R_x + R_y < H(\bar{x},\bar{y})} D(p_{\bar{x},\bar{y}} \| p_{xy}) \right\} \tag{3}$$

there does not exists a randomized encoder-decoder mapping as defined in Definition 1 such that (1) is satisfied.

In both bounds $(\bar{x}, \bar{y})$ are dummy random variables with joint distribution $p_{\bar{x},\bar{y}}$.

Remark: As long as (R_x, R_y) is in the interior of the achievable region, i.e., $R_x > H(x|y)$, $R_y > H(y|x)$ and $R_x + R_y > H(x, y)$ then the lower-bound (2) is positive. The achievable region is illustrated in Fig 2. As shown in [8], the upper and lower bounds (3) and (2) match when the rate pair (R_x, R_y) is achievable and close to the boundary of the region. This is analogous to the high rate regime in channel coding where the random coding bound (analogous to (2)) and the sphere packing bound (analogous to (3)) agree.

Theorem 1 can also be used to generate bounds on the exponent for source coding with decoder side information (i.e., $\mathbf{y}$ observed at the decoder), and for source coding without side information (i.e., $\mathbf{y}$ is a constant). These corollaries will prove useful as a basis for comparison as we build up to the complete solution for streaming Slepian-Wolf coding.

Corollary 1: (Source coding with decoder side information) Consider a Slepian-Wolf problem where $\mathbf{y}$ is known by the decoder. Given a rate R_x such that $R_x > H(x|y)$, then for all

$$E < \min_{\bar{x},\bar{y}} D(p_{\bar{x},\bar{y}} \| p_{xy}) + |R_x - H(\bar{x}|\bar{y})|^+, \tag{4}$$

there exists a family of randomized encoder-decoder mappings as defined in Definition 1 such that (1) is satisfied.

The proof of Corollary 1 follows from Theorem 1 by letting R_y be arbitrarily large. Similarly, by letting $\mathbf{y}$ be deterministic so that $H(x|y) = H(x)$ and $H(y) = 0$, we get the following random-coding bound for the point-to-point case of a single source $\mathbf{x}$.

Corollary 2: (point-to-point) Consider a Slepian-Wolf problem where $\mathbf{y}$ is deterministic, i.e., $\mathbf{y} = \mathbf{y}$. Given a rate R_x such that $R_x > H(x)$, for all

$$E < \min_{\bar{x}} D(p_{\bar{x}} \| p_x) + |R_x - H(\bar{x})|^+ = E_x(R_x) \tag{5}$$

there exists a family of randomized encoder-decoder triplet as defined in Definition 1 such that (1) is satisfied.

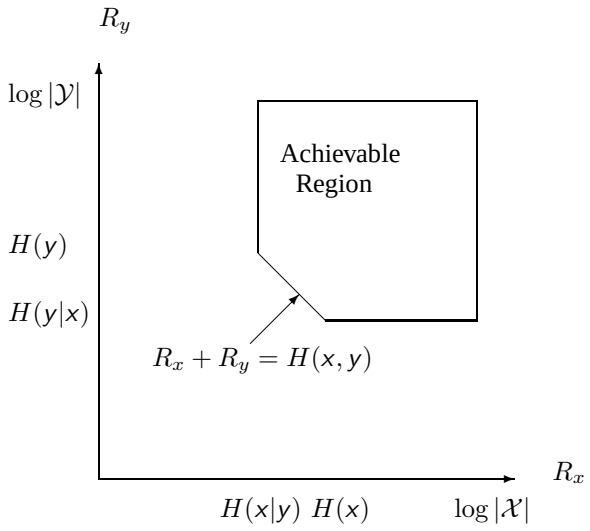

Fig. 2. Achievable region for Slepian-Wolf source coding

B. Sequential Distributed Source Coding

We now state our main results for streaming encoding, and contrast them with the block-coding results of the last section. To begin, we define a streaming encoder.

Definition 2: A randomized sequential encoder-decoder triplet $\mathcal{E}^x, \mathcal{E}^y, \mathcal{D}$ is a sequence of mappings, $\{\mathcal{E}_j^x\}, j = 1, 2, ...$, $\{\mathcal{E}_j^y\}, j = 1, 2, ...$ and $\{\mathcal{D}_j\}, j = 1, 2, ...$:

$$\begin{array}{lll} \mathcal{E}_j^x & : & \mathcal{X}^j \longrightarrow \{0,1\}^{R_x}, \quad \text{e.g.,} \quad \mathcal{E}_j^x(x^j) = a_{(j-1)R_x+1}^{jR_x}, \\ \mathcal{E}_j^y & : & \mathcal{Y}^j \longrightarrow \{0,1\}^{R_y}, \quad \text{e.g.,} \quad \mathcal{E}_j^y(y^j) = b_{(j-1)R_y+1}^{jR_y}. \end{array} \tag{6}$$

Common randomness, shared between encoders and decoder, is assumed. This allows us to randomize the mappings independently of the source sequence.

In this paper, the sequential encoding maps will always work by assigning random "parity bits" in a causal manner to the observed source sequence. That is, the R_x (or R_y) bits generated at each time in (6), are iid Bernoulli-(0.5).[12] Since parity bits are assigned causally, if two source sequences share the same length-l prefix, then their first lR_x parity bits must match. Subsequent parities are drawn independently. Such a sequential coding strategy is the source-coding parallel to tree and convolutional codes used for channel coding [10]. In fact, we call these "parity bits" as they can be generated using an infinite constraint-length time-varying random convolutional code.

Definition 3: The decoder mapping

$$\mathcal{D}_j : \{0,1\}^{jR_x} \times \{0,1\}^{jR_y} \longrightarrow \mathcal{X}^j \times \mathcal{Y}^j$$
$$\mathcal{D}_j(a^{jR_x}, b^{jR_y}) = (\hat{x}_1^j(j), \hat{y}_1^j(j))$$

At each time j the decoder $\mathcal{D}_j$ outputs estimates of all the source symbols that have entered the encoder by time j.

Remark: While we state Definition 2 only for Slepian-Wolf coding, it immediately specializes to source coding with decoder side information (dropping the $\mathcal{E}_y$ and revealing y^n to the decoder), and source coding without side information (dropping the $\mathcal{E}_y$). We present results for both these situations as well.

In this paper we study two error probabilities. We define the pair of source estimates at time n as $(\hat{x}^n, \hat{y}^n) = \mathcal{D}_n(\prod_{j=1}^n \mathcal{E}_j^x, \prod_{j=1}^n \mathcal{E}_j^y)$, where $\prod_{j=1}^n \mathcal{E}_j^x$ indicates the full nR_x bit stream from encoder x up to time n. We use $(\hat{x}^{n-\Delta}, \hat{y}^{n-\Delta})$ to indicate the first $n-\Delta$ symbols of each estimate, where for conciseness of notation both the estimate time, n, and the decoding delay, Δ, are indicated in the superscript. With these definitions the two error probabilities we study are

$$\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \text{ and } \Pr[\hat{y}^{n-\Delta} \neq y^{n-\Delta}].$$

A pair of exponents $E_x > 0$ and $E_y > 0$ is said to be achievable if there exists a family of rate-(R_x, R_y) encoders and decoders $\{(\mathcal{E}_j^x, \mathcal{E}_j^y, \mathcal{D}_j)\}$ such that

$$\lim_{\Delta\to\infty} \lim_{n\to\infty} -\frac{1}{\Delta} \log \Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \geq E_x \tag{7}$$

$$\lim_{\Delta\to\infty} \lim_{n\to\infty} -\frac{1}{\Delta} \log \Pr[\hat{y}^{n-\Delta} \neq y^{n-\Delta}] \geq E_y \tag{8}$$

Remarks: In contrast to (1) the error exponent we look at is in the delay, Δ, rather than total observation time, n. The order of the limits is important since the total time-period n is allowed to go to infinity faster than the delay Δ. While the definitions of (7)–(8) and of (1) are asymptotic in nature, the results hold for finite block-lengths and delays as well. Finally, we note that while in (1) the error exponent of a joint error event on either $\mathbf{x}$ or $\mathbf{y}$ is considered, we provide a refined analysis specifying potentially different exponents on either decision. The results for joint errors are found by taking the minimum of the individual exponents, i.e.,

$$\lim_{\Delta\to\infty} \lim_{n\to\infty} -\frac{1}{\Delta} \log \Pr[(\hat{x}^{n-\Delta}, \hat{y}^{n-\Delta}) \neq (x^{n-\Delta}, y^{n-\Delta})] \geq \min\{E_x, E_y\}.$$

C. Streaming source coding

Our first results concern streaming coding in the point-to-point setting. The first theorem we state gives random coding error exponents for maximum likelihood decoding where the source statistics are known, and the second exponents for universal decoding, where they are not.

[12]We assume that R_x and R_y are integer. To justify this assumption note that we can always group sets of α successive symbols into super-symbols. These larger symbols can be encoded at an average rate αR_x. Generally, if we group α symbols together, and transmit β bits per super-symbol, we can realize an average rate α/β, i.e., a rational rate. If desired, non-integer average rates are easily implemented by a time-varying transmission rate. For example, say we want to implement an average encoding rate of $5/4$ bits per source symbol. Say we generate one new parity bit per symbol for each symbol observed except for the fourth symbol, eighth symbol, etc, when we generate two. The average encoding rate is $5/4$. As long as the decoding delay Δ we target is long enough so that the decoder received an "average" number of encoded bits – δR_x – before we must make an estimate (e.g., if $\Delta \gg 1/R_x$), these small-scale issues even out. In particular, they do not effect the exponents.

Theorem 2: Given a rate $R_x > H(p_{\mathsf{x}})$, there exists a randomized streaming encoder and maximum likelihood decoder pair (per Definition 2) such that for all $E < E_{ML}(R_x)$ there is a constant $K > 0$ such that $\Pr[\hat{\mathsf{x}}^{n-\Delta} \neq \mathsf{x}^{n-\Delta}] \leq K \exp\{-\Delta E_{ML}(R_x)\}$ for all $n, \Delta \geq 0$ where

$$E_{ML}(R_x) = \sup_{0 \leq \rho \leq 1} \rho R_x - (1+\rho) \log \left(\sum_x p_{\mathsf{x}}(x)^{\frac{1}{1+\rho}} \right). \tag{9}$$

Theorem 3: Given a rate $R_x > H(p_{\mathsf{x}})$, there exists a randomized streaming encoder and universal decoder pair (per Definition 2) such that for all $E < E_{UN}(R_x)$ there is a constant $K > 0$ such that $\Pr[\hat{\mathsf{x}}^{n-\Delta} \neq \mathsf{x}^{n-\Delta}] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{UN}(R_x) = \inf_q D(q \| p_{\mathsf{x}}) + |R_x - H(q)|^+, \tag{10}$$

where q is an arbitrary probability distribution on $\mathcal{X}$ and where $|z|^+ = z$ if $z \geq 0$ and $|z|^+ = 0$ if $z < 0$.

Remark: The error exponents of Theorems 2 and 3 both equal their respective random block-coding exponents for ML and universal decoders. For example, compare (10) with (5). The main difference in the formulation is that the error probability decays with delay Δ rather than block length N. Furthermore, it is known that (9) and (10) are equal — see [8] exercise 13 on page 44. Such equality is required by the formal definition of a universal scheme, i.e., for the same source statistics and coding rates, the universal decoder should asymptotically achieve the same error exponent as the maximum likelihood decoder. See [11] for a detailed discussion of universal versus maximum likelihood decoding in the context of channel coding.

D. Streaming distributed source coding with decoder side information

This section summarizes our results for distributed streaming source coding when the side information is observed at the decoder, but not the encoder:

Theorem 4: Given a rate $R_x > H(\mathsf{x}|\mathsf{y})$, there exists a randomized encoder decoder pair (per Definition 2) such that for all $E < E_{ML,SI}(R_x)$ there is a constant $K > 0$ such that $\Pr[\hat{\mathsf{x}}^{n-\Delta} \neq \mathsf{x}^{n-\Delta}] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{ML,SI}(R_x) = \sup_{0 \leq \rho \leq 1} \rho R_x - \log \Big[\sum_y \Big[\sum_x p_{\mathsf{xy}}(x,y)^{\frac{1}{1+\rho}} \Big]^{1+\rho} \Big]. \tag{11}$$

Theorem 5: Given a rate $R_x > H(\mathsf{x}|\mathsf{y})$, there exists a randomized encoder decoder pair (per Definition 2) such that for all $E < E_{UN,SI}(R_x)$ there is a constant $K > 0$ such that $\Pr[\hat{\mathsf{x}}^{n-\Delta} \neq \mathsf{x}^{n-\Delta}] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{UN,SI}(R_x) = \inf_{\tilde{\mathsf{x}}, \tilde{\mathsf{y}}} D(p_{\tilde{\mathsf{x}}, \tilde{\mathsf{y}}} \| p_{\mathsf{xy}}) + |R_x - H(\tilde{\mathsf{x}}|\tilde{\mathsf{y}})|^+, \tag{12}$$

and $(\tilde{\mathsf{x}}, \tilde{\mathsf{y}})$ are random variables with joint distribution $p_{\tilde{\mathsf{x}}, \tilde{\mathsf{y}}}$, $H(\tilde{\mathsf{x}}|\tilde{\mathsf{y}})$ is their conditional entropy, and where $|z|^+ = z$ if $z \geq 0$ and $|z|^+ = 0$ if $z < 0$.

Remark: Similar to the point-to-point case, the error exponents of Theorems 4 and 5 both equal their respective random block-coding exponents. For example, compare (12) with (4). Similarly, (11) and (12) can be shown to be equal.

E. Streaming Slepian-Wolf coding

In contrast to streaming point-to-point coding and streaming source coding with decoder side information, the general case of streaming Slepian-Wolf coding with two distributed encoders results in error exponents that differ from their block coding counterparts. In the streaming setting, fundamentally different error events dominate as compared to the block setting.

Theorem 6: Let (R_x, R_y) be a rate pair such that $R_x > H(\mathsf{x}|\mathsf{y})$, $R_y > H(\mathsf{y}|\mathsf{x})$, $R_x + R_y > H(\mathsf{x}, \mathsf{y})$. Then, there exists a randomized encoder pair and maximum likelihood decoder triplet (per Definition 2) that satisfies the following three decoding criteria.

(i) For all $E < E_{ML,SW,x}(R_x, R_y)$, there is a constant $K > 0$ such that $\Pr[\hat{\mathsf{x}}^{n-\Delta} \neq \mathsf{x}^{n-\Delta}] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{ML,SW,x}(R_x, R_y) = \min \left\{ \inf_{\gamma \in [0,1]} E_x^{ML}(R_x, R_y, \gamma), \inf_{\gamma \in [0,1]} \frac{1}{1-\gamma} E_y^{ML}(R_x, R_y, \gamma) \right\}.$$

(ii) For all $E < E_{ML,SW,y}(R_x, R_y)$ there is a constant $K > 0$ such that $\Pr[\hat{y}^{n-\Delta} \neq y^{n-\Delta}] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{ML,SW,y}(R_x, R_y) = \min \left\{ \inf_{\gamma \in [0,1]} \frac{1}{1-\gamma} E_x^{ML}(R_x, R_y, \gamma), \inf_{\gamma \in [0,1]} E_y^{ML}(R_x, R_y, \gamma) \right\}.$$

(iii) For all $E < E_{ML,SW,xy}(R_x, R_y)$ there is a constant $K > 0$ such that $\Pr[(\hat{x}^{n-\Delta}, \hat{y}^{n-\Delta}) \neq (x^{n-\Delta}, y^{n-\Delta})] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{ML,SW,xy}(R_x, R_y) = \min \left\{ \inf_{\gamma \in [0,1]} E_x^{ML}(R_x, R_y, \gamma), \inf_{\gamma \in [0,1]} E_y^{ML}(R_x, R_y, \gamma) \right\}.$$

In definitions (i)–(iii),

$$\begin{aligned} E_x^{ML}(R_x, R_y, \gamma) &= \sup_{\rho \in [0,1]} [\gamma E_{x|y}(R_x, \rho) + (1-\gamma) E_{xy}(R_x, R_y, \rho)] \\ E_y^{ML}(R_x, R_y, \gamma) &= \sup_{\rho \in [0,1]} [\gamma E_{y|x}(R_x, \rho) + (1-\gamma) E_{xy}(R_x, R_y, \rho)] \end{aligned} \tag{13}$$

and

$$\begin{aligned} E_{xy}(R_x, R_y, \rho) &= \rho(R_x + R_y) - \log \left[\sum_{x,y} p_{xy}(x,y)^{\frac{1}{1+\rho}} \right]^{1+\rho} \\ E_{x|y}(R_x, \rho) &= \rho R_x - \log \left[\sum_y \left[\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}} \right]^{1+\rho} \right] \\ E_{y|x}(R_y, \rho) &= \rho R_y - \log \left[\sum_x \left[\sum_y p_{xy}(x,y)^{\frac{1}{1+\rho}} \right]^{1+\rho} \right] \end{aligned} \tag{14}$$

Theorem 7: Let (R_x, R_y) be a rate pair such that $R_x > H(x|y)$, $R_y > H(y|x)$, $R_x + R_y > H(x,y)$. Then, there exists a randomized encoder pair and universal decoder triplet (per Definition 2) that satisfies the following three decoding criteria.

(i) For all $E < E_{UN,SW,x}(R_x, R_y)$, there is a constant $K > 0$ such that $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{UN,SW,x}(R_x, R_y) = \min \left\{ \inf_{\gamma \in [0,1]} E_x^{UN}(R_x, R_y, \gamma), \inf_{\gamma \in [0,1]} \frac{1}{1-\gamma} E_y^{UN}(R_x, R_y, \gamma) \right\}. \tag{15}$$

(ii) For all $E < E_{UN,SW,y}(R_x, R_y)$, there is a constant $K > 0$ such that $\Pr[\hat{y}^{n-\Delta} \neq y^{n-\Delta}] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{UN,SW,y}(R_x, R_y) = \min \left\{ \inf_{\gamma \in [0,1]} \frac{1}{1-\gamma} E_x^{UN}(R_x, R_y, \gamma), \inf_{\gamma \in [0,1]} E_y^{UN}(R_x, R_y, \gamma) \right\}. \tag{16}$$

(iii) For all $E < E_{UN,SW,xy}(R_x, R_y)$, there is a constant $K > 0$ such that $\Pr[(\hat{x}^{n-\Delta}, \hat{x}^{n-\Delta}) \neq (x^{n-\Delta}, y^{n-\Delta})] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{UN,SW,xy}(R_x, R_y) = \min \left\{ \inf_{\gamma \in [0,1]} E_x^{UN}(R_x, R_y, \gamma), \inf_{\gamma \in [0,1]} E_y^{UN}(R_x, R_y, \gamma) \right\}. \tag{17}$$

In definitions (i)–(iii),

$$\begin{aligned} &E_x^{UN}(R_x, R_y, \gamma) \\ &= \inf_{\tilde{x}, \tilde{y}, \bar{x}, \bar{y}} \gamma D(p_{\tilde{x},\tilde{y}} \| p_{xy}) + (1-\gamma) D(p_{\bar{x},\bar{y}} \| p_{xy}) + |\gamma [R_x - H(\tilde{x}|\tilde{y})] + (1-\gamma)[R_x + R_y - H(\bar{x}, \bar{y})]|^+ \\ &E_y^{UN}(R_x, R_y, \gamma) \\ &= \inf_{\tilde{x}, \tilde{y}, \bar{x}, \bar{y}} \gamma D(p_{\tilde{x},\tilde{y}} \| p_{xy}) + (1-\gamma) D(p_{\bar{x},\bar{y}} \| p_{xy}) + |\gamma [R_y - H(\tilde{y}|\tilde{x})] + (1-\gamma)[R_x + R_y - H(\bar{x}, \bar{y})]|^+ \end{aligned} \tag{18}$$

where the random variables $(\tilde{x}, \tilde{y})$ and $(\bar{x}, \bar{y})$ have joint distributions $p_{\tilde{x},\tilde{y}}$ and $p_{\bar{x},\bar{y}}$, respectively. The function $|z|^+ = z$ if $z \geq 0$ and $|z|^+ = 0$ if $z < 0$.

Remark: Definitions (i) and (ii) in Theorems 6 and 7 concern individual decoding error events which might be useful in applications where the $\mathbf{x}$ and $\mathbf{y}$ streams are decoded jointly, but utilized individually. The more standard joint error event is given by (iii).

Remark: We can compare the joint error event for block and streaming Slepian-Wolf coding, c.f. (17) with (2). The streaming exponent differs by the extra parameter γ that must be minimized over. If the minimizing $\gamma = 1$, then the block and streaming exponents are the same. The minimization over γ results from a fundamental difference in the types of error-causing events that can occur in streaming Slepian-Wolf as compared to block Slepian-Wolf.

Remark: The error exponents of maximum likelihood and universal decoding in Theorems 6 and 7 are the same. However, because there are new classes of error events possible in streaming, this needs proof. The equivalence is summarized in the following theorem.

Theorem 8: Let (R_x, R_x) be a rate pair such that $R_x > H(x|y)$, $R_y > H(y|x)$, and $R_x + R_y > H(x, y)$. Then,

$$E_{ML,SW,x}(R_x, R_y) = E_{UN,SW,x}(R_x, R_y), \tag{19}$$

and

$$E_{ML,SW,x}(R_x, R_y) = E_{UN,SW,x}(R_x, R_y). \tag{20}$$

Theorem 8 follows directly from the following lemma, shown in the appendix.

Lemma 1: For all $\gamma \in [0, 1]$

$$E_x^{ML}(R_x, R_y, \gamma) = E_x^{UN}(R_x, R_y, \gamma), \tag{21}$$

and

$$E_y^{ML}(R_x, R_y, \gamma) = E_y^{UN}(R_x, R_y, \gamma). \tag{22}$$

.

Remark: This theorem allows us to simplify notation. For example, we can define $E_x(R_x, R_y, \gamma)$ as $E_x(R_x, R_y, \gamma) = E_x^{ML}(R_x, R_y, \gamma) = E_x^{UN}(R_x, R_y, \gamma)$, and can similarly define $E_y(R_x, R_y, \gamma)$. Further, since the ML and universal exponents are the same for the whole rate region we can define $E_{SW,x}(R_x, R_y)$ as $E_{SW,x}(R_x, R_y) = E_{ML,SW,x}(R_x, R_y) = E_{UN,SW,x}(R_x, R_y)$, and can similarly define $E_{SW,y}(R_x, R_y)$.

IV. Numerical Results

To build insight into the differences between the sequential error exponents of Theorem 2 - 8 and block-coding error exponents, we give some examples of the exponents for binary sources.

For the point-to-point case, the error exponents of random sequential and block source coding are identical everywhere in the achievable rate region as can be seen by comparing Theorem 3 and Corollary 2. The same is true for source coding with decoder side information (cf. Theorem 5 and Corollary 1). For distributed Slepian-Wolf source coding however, the sequential and block error exponents can be different. The reason for the discrepancy is that a new type of error event can be dominant in Slepian-Wolf source coding. This is reflected in Theorem 6 by the minimization over γ. Example 2 illustrates the impact of this γ term.

For Slepian-Wolf source coding at very high rates, where $R_x > H(x)$, the decoder can ignore any information from encoder y and still decode x with with a positive error exponent. However, the decoder could also choose to decode source x and y jointly. Fig 6.a and 6.b illustrate that joint decoding may or surprisingly *may not* help decoding source x. This is seen by comparing the error exponent when the decoder ignores the side information from encoder y (the dotted curves) to the joint error exponent (the lower solid curves). It seems that when the rate for source y is low, atypical behaviors of source y can cause joint decoding errors that end up corrupting x estimates. This holds for both block and sequential coding.

A. Example 1: symmetric source with uniform marginals

Consider a symmetric source where $|\mathcal{X}| = |\mathcal{Y}| = 2$, $p_{xy}(0, 0) = 0.45$, $p_{xy}(0, 1) = p_{xy}(1, 0) = 0.05$ and $p_{xy}(1, 1) = 0.45$. This is a marginally-uniform source: x is Bernoulli(1/2), y is the output from a BSC with input x, thus y is Bernoulli(1/2) as well. For this source $H(x) = H(y) = \log(2)$, $H(x|y) = H(y|x) = 0.32$, $H(x, y) = 1.02$. The achievable rate region is the triangle shown in Figure(3).

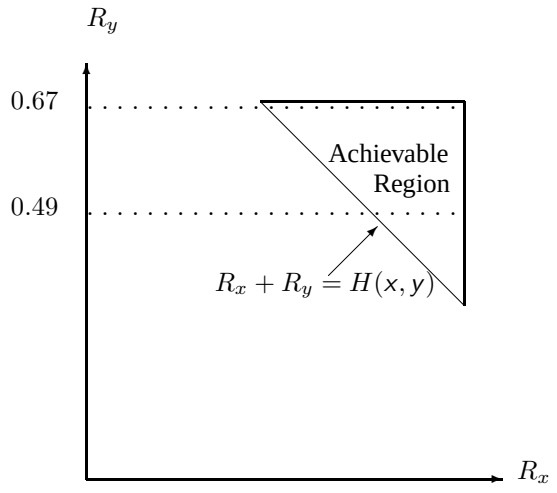

Fig. 3. Rate region for the example 1 source, we focus on the error exponent on source x for fixed encoder y rates: $R_y = 0.49$ and $R_y = 0.67$

For this source, as will be shown later, the dominant sequential error event is on the diagonal line in Fig 9. This is to say that:

$$E_{SW,x}(R_x, R_y) = E_{SW,x}^{BLOCK}(R_x, R_y) = E_x^{ML}(R_x, R_y, 0) = \sup_{\rho\in[0,1]} [E_{xy}(R_x, R_y, \rho)]. \tag{23}$$

Where $E_{SW,x}^{BLOCK}(R_x, R_y) = \min\{E_x^{ML}(R_x, R_y, 0), E_x^{ML}(R_x, R_y, 1)\}$ as shown in [9].
Similarly for source y:

$$E_{SW,y}(R_x, R_y) = E_{SW,y}^{BLOCK}(R_x, R_y) = E_y^{ML}(R_x, R_y, 0) = \sup_{\rho\in[0,1]} [E_{xy}(R_x, R_y, \rho)]. \tag{24}$$

We first show that for this source $\forall \rho \geq 0$, $E_{x|y}(R_x, \rho) \geq E_{xy}(R_x, R_y, \rho)$. By definition:

$$\begin{aligned}
E_{x|y}(R_x,\rho) - E_{xy}(R_x,R_y,\rho) &= \rho R_x - \log\Big[\sum_y \Big[\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big] \\
&\quad -\Big(\rho(R_x+R_y) - \log\Big[\sum_{x,y} p_{xy}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big) \\
&= -\rho R_y - \log\Big[2\Big[\sum_x p_{xy}(x,0)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big] + \log\Big[2\sum_x p_{xy}(x,0)^{\frac{1}{1+\rho}}\Big]^{1+\rho} \\
&= -\rho R_y - \log\Big[2\Big] + \log\Big[2\Big]^{1+\rho} \\
&= \rho(\log[2] - R_y) \\
&\geq 0
\end{aligned}$$

The last inequality is true because we only consider the problem when $R_y \leq \log|\mathcal{Y}|$. Otherwise, y is better viewed as perfectly known side-information. Now

$$\begin{aligned}
E_x^{ML}(R_x,R_y,\gamma) &= \sup_{\rho\in[0,1]} [\gamma E_{x|y}(R_x,\rho) + (1-\gamma)E_{xy}(R_x,R_y,\rho)] \\
&\geq \sup_{\rho\in[0,1]} [E_{xy}(R_x,R_y,\rho)] \\
&= E_x^{ML}(R_x,R_y,0)
\end{aligned}$$

Similarly $E_y^{ML}(R_x, R_y, \gamma) \geq E_y^{ML}(R_x, R_y, 0) = E_x^{ML}(R_x, R_y, 0)$. Finally,

$$\begin{aligned} E_{SW,x}(R_x, R_y) &= \min\left\{ \inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma), \inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x, R_y, \gamma) \right\} \\ &= E_x^{ML}(R_x, R_y, 0) \end{aligned}$$

Particularly $E_x(R_x, R_y, 1) \geq E_x(R_x, R_y, 0)$, so

$$\begin{aligned} E_{SW,x}^{BLOCK}(R_x, R_y) &= \min\{E_x^{ML}(R_x, R_y, 0), E_x^{ML}(R_x, R_y, 1)\} \\ &= E_x^{ML}(R_x, R_y, 0) \end{aligned}$$

The same proof holds for source y.

In Fig 4 we plot the joint sequential/block coding error exponents $E_{SW,x}(R_x, R_y) = E_{SW,x}^{BLOCK}(R_x, R_y)$, the error exponents are positive iff $R_x > H(xy) - R_y = 1.02 - R_y$.

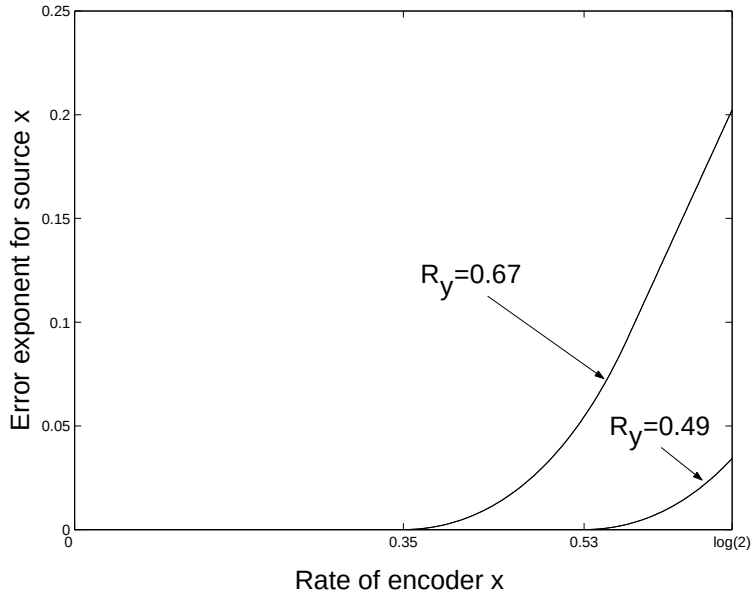

Fig. 4. Error exponents plot: $E_{SW,x}(R_x, R_y)$ plotted for $R_y = 0.49$ and $R_y = 0.67$
$E_{SW,x}(R_x, R_y) = E_{SW,x}^{BLOCK}(R_x, R_y) = E_{SW,y}(R_x, R_y) = E_{SW,y}^{BLOCK}(R_x, R_y)$ and $E_x(R_x) = 0$

B. Example 2: non-symmetric source

Consider a non-symmetric source where $|\mathcal{X}| = |\mathcal{Y}| = 2$, $p_{xy}(0,0) = 0.1$, $p_{xy}(0,1) = p_{xy}(1,0) = 0.05$ and $p_{xy}(1,1) = 0.8$. For this source $H(x) = H(y) = 0.42$, $H(x|y) = H(y|x) = 0.29$ and $H(x,y) = 0.71$. The achievable rate region is shown in Fig 5. In Fig 6.a, 6.b, 6.c and 6.d, we compare the joint sequential error exponent $E_{SW,x}(R_x, R_y)$ the joint block coding error exponent $E_{SW,x}^{BLOCK}(R_x, R_y) = \min\{E_x(R_x, R_y, 0), E_x(R_x, R_y, 1)\}$ as shown in [9] and the individual error exponent for source X, $E_x(R_x)$ as shown in Corollary 2. Notice that $E_x(R_x) > 0$ only if $R_x > H(x)$. In Fig 7, we compare the sequential error exponent for source y: $E_{SW,y}(R_x, R_y)$ and the block coding error exponent for source y: $E_{SW,y}^{BLOCK}(R_x, R_y) = \min\{E_y(R_x, R_y, 0), E_y(R_x, R_y, 1)\}$ and $E_y(R_y)$ which is a constant since we fix R_y.

For $R_y = 0.35$ as shown in Fig 6.a.b and 7.a.b, the difference between the block coding and sequential coding error exponents is very small for both source x and y. More interestingly, as shown in Fig 6.a, because the rate of source y is low, i.e. it is more likely to get a decoding error due to the atypical behavior of source y. So as R_x

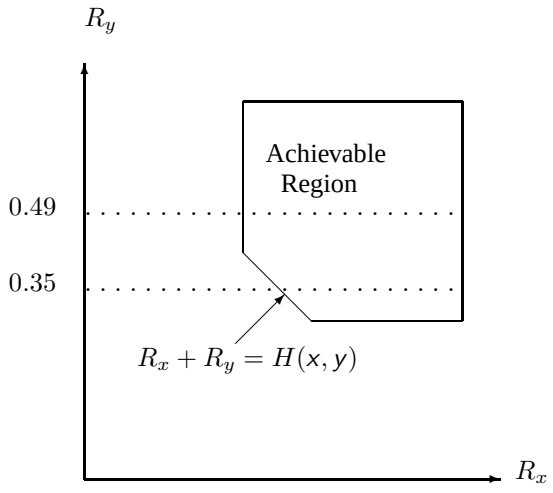

Fig. 5. Rate region for the example 2 source, we focus on the error exponent on source x for fixed encoder y rates: $R_y = 0.35$ and $R_y = 0.49$

increases, it is sometimes better to ignore source y and decode x individually. This is evident as the dotted curve is above the solid curves.

For $R_y = 0.49$ as shown in Fig 6.c.d and 7.c.d, since the rate for source y is high enough, source y can be decoded with a positive error exponent individually as shown in Fig 7.c. But as the rate of source x increases, joint decoding gives a better error exponent. When R_x is very high, then we observe the saturation of the error exponent on y as if source x is known perfectly to the decoder! This is illustrated by the flat part of the solid curves in Fig 7.c.

V. Streaming point-to-point coding via sequential random binning

In this section we prove Theorems 2 and 3. While the emphasis of the paper is on distributed source coding, the basic causal random binning ideas and analysis techniques can be more easily developed in the point-to-point context.

A. Maximum-likelihood decoding

To show Theorems 2 and 3, we first develop the common core of the proof in the context of ML decoding. The proof strategy is as follows. A decoding error can only occur if there is some spurious source sequence $\tilde{x}^n$ that satisfies three conditions: (i) it must be in the same bin (share the same parities) as x^n, i.e., $\tilde{x}^n \in \mathcal{B}_x(x^n)$, (ii) it must be more likely than the true sequence, i.e., $p_{\mathbf{x}}(\tilde{x}^n) > p_{\mathbf{x}}(x^n)$, and (iii) $\tilde{x}_l \neq x_l$ for some $l \leq n - \Delta$.

The error probability is

$$\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] = \sum_{x^n} \Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta} | x^n = x^n] p_{\mathbf{x}}(x^n) \tag{25}$$

$$= \sum_{x^n} \sum_{l=1}^{n-\Delta} \Pr\left[\exists\, \tilde{x}^n \in \mathcal{B}_x(x^n) \cap \mathcal{F}_n(l, x^n) \text{ s.t. } p_{\mathbf{x}}(\tilde{x}^n) \geq p_{\mathbf{x}}(x^n)\right] p_{\mathbf{x}}(x^n) \tag{26}$$

$$= \sum_{l=1}^{n-\Delta} \left\{ \sum_{x^n} \Pr\left[\exists\, \tilde{x}^n \in \mathcal{B}_x(x^n) \cap \mathcal{F}_n(l, x^n) \text{ s.t. } p_{\mathbf{x}}(\tilde{x}^n) \geq p_{\mathbf{x}}(x^n)\right] p_{\mathbf{x}}(x^n) \right\}$$

$$= \sum_{l=1}^{n-\Delta} p_n(l). \tag{27}$$

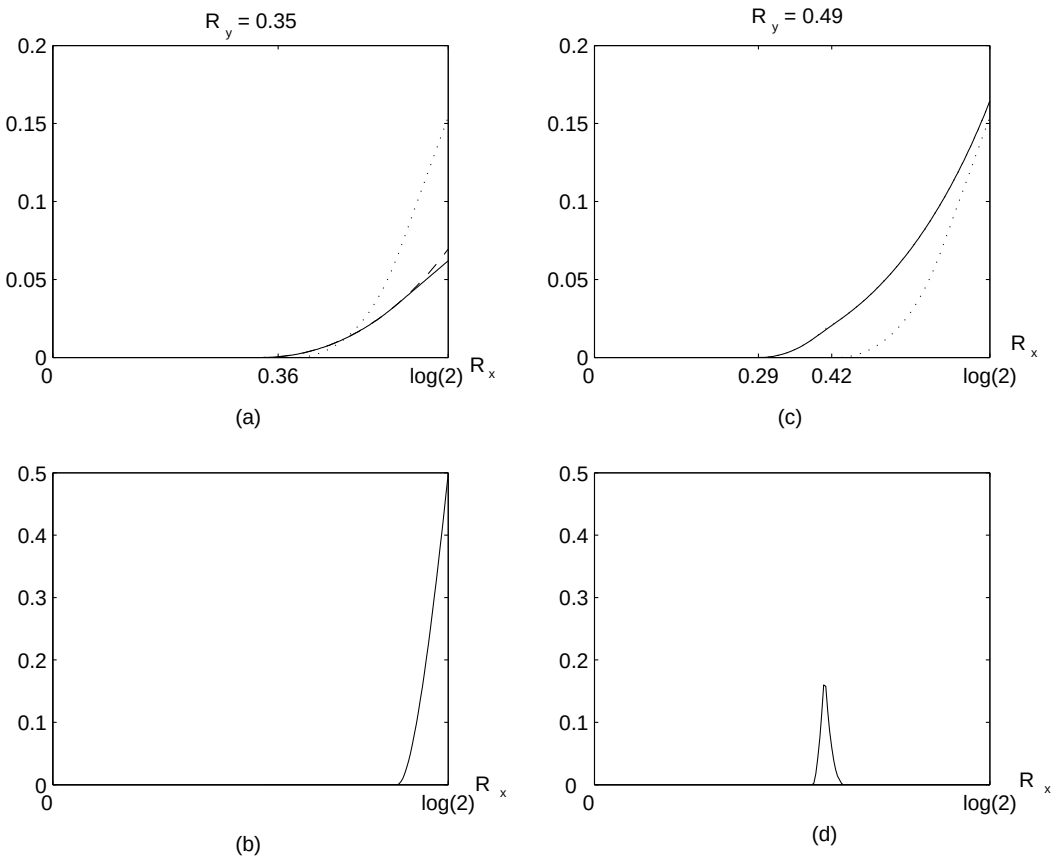

Fig. 6. Error exponents plot for source x for fixed R_y as R_x varies:
$R_y = 0.35$:
(a) Solid curve: $E_{SW,x}(R_x, R_y)$, dashed curve $E_{SW,x}^{BLOCK}(R_x, R_y)$ and dotted curve: $E_x(R_x)$, notice that $E_{SW,x}(R_x, R_y) \leq E_{SW,x}^{BLOCK}(R_x, R_y)$ but the difference is small.
(b) $10\log_{10}(\frac{E_{SW,x}^{BLOCK}(R_x,R_y)}{E_{SW,x}(R_x,R_y)})$. This shows the difference is there at high rates.
$R_y = 0.49$:
(c) Solid curve $E_{SW,x}(R_x, R_y)$, dashed curve $E_{SW,x}^{BLOCK}(R_x, R_y)$ and dotted curve: $E_x(R_x)$, again $E_{SW,x}(R_x, R_y) \leq E_{SW,x}^{BLOCK}(R_x, R_y)$ but the difference is extremely small.
(d) $10\log_{10}(\frac{E_{SW,x}^{BLOCK}(R_x,R_y)}{E_{SW,x}(R_x,R_y)})$. This shows the difference is there at intermediate low rates.

After conditioning on the realized source sequence in (25), the remaining randomness is only in the binning. In (26) we decompose the error event into a number of mutually exclusive events (see Fig 8) by partitioning all source sequences $\tilde{x}^n$ into sets $\mathcal{F}_n(l, x^n)$ defined by the time l of the first sample in which they differ from the realized source x^n,

$$\mathcal{F}_n(l, x^n) = \{\tilde{x}^n \in \mathcal{X}^n | \tilde{x}^{l-1} = x^{l-1}, \tilde{x}_l \neq x_l\}, \tag{28}$$

and define $\mathcal{F}_n(n+1, x^n) = \{x^n\}$. Finally, in (27) we define

$$p_n(l) = \sum_{x^n} \Pr\left[\exists\, \tilde{x}^n \in \mathcal{B}_x(x^n) \cap \mathcal{F}_n(l, x^n) \text{ s.t. } p_{\mathbf{x}}(\tilde{x}^n) \geq p_{\mathbf{x}}(x^n)\right] p_{\mathbf{x}}(x^n). \tag{29}$$

We now upper bound $p_n(l)$ using a Chernoff bound argument similar to [9].
Lemma 2: $p_n(l) \leq \exp\{-(n-l+1)E_{ML}(R_x)\}$.

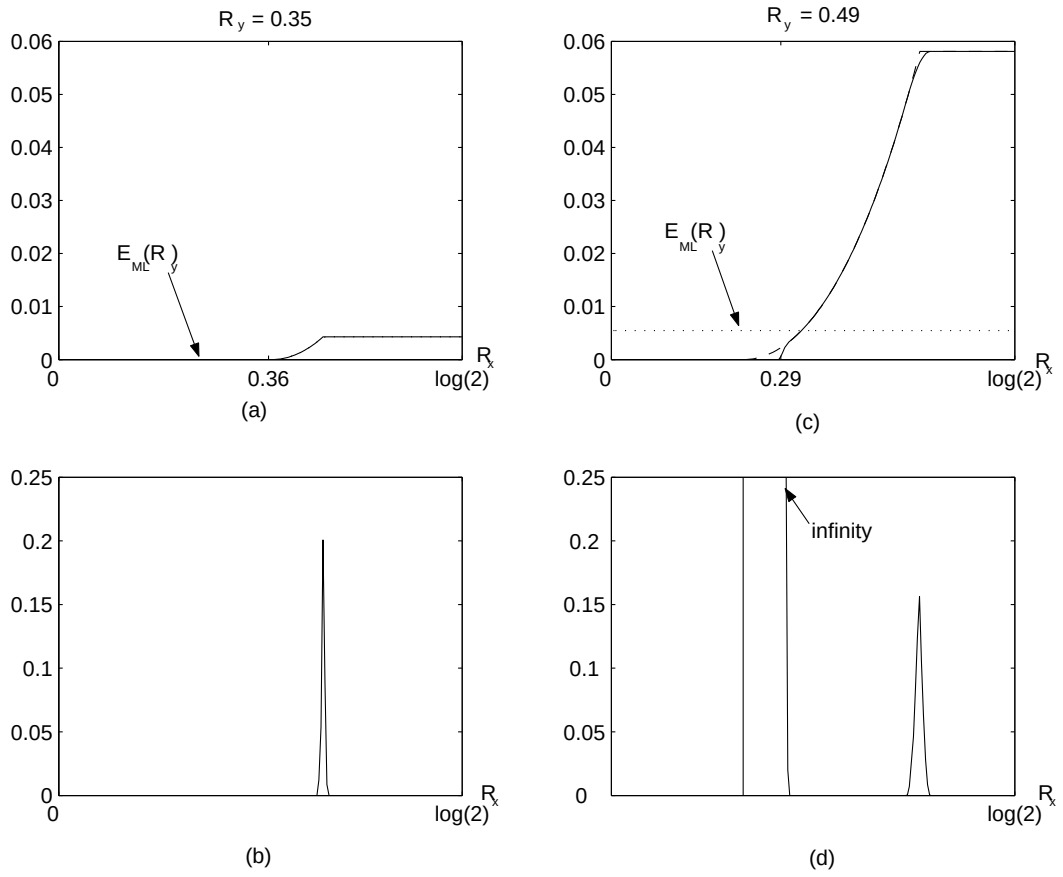

Fig. 7. Error exponents plot for source y for fixed R_y as R_x varies:
$R_y = 0.35$:
(a) Solid curve: $E_{SW,y}(R_x, R_y)$ and dashed curve $E_{SW,y}^{BLOCK}(R_x, R_y)$, $E_{SW,y}(R_x, R_y) \leq E_{SW,y}^{BLOCK}(R_x, R_y)$, the difference is extremely small. $E_y(R_y)$ is 0 because $R_y = 0.35 < H(y)$. (b) $10 \log_{10}(\frac{E_{SW,y}^{BLOCK}(R_x, R_y)}{E_{SW,y}(R_x, R_y)})$. This shows the two exponents are not identical everywhere.
$R_y = 0.49$:
(c) Solid curves: $E_{SW,y}(R_x, R_y)$, dashed curve $E_{SW,y}^{BLOCK}(R_x, R_y)$ and $E_{SW,y}(R_x, R_y) \leq E_{SW,y}^{BLOCK}(R_x, R_y)$ and $E_y(R_y)$ is constant shown in a dotted line.
(d) $10 \log_{10}(\frac{E_{SW,y}^{BLOCK}(R_x, R_y)}{E_{SW,y}(R_x, R_y)})$. Notice how the gap goes to infinity when we leave the Slepian-Wolf region.

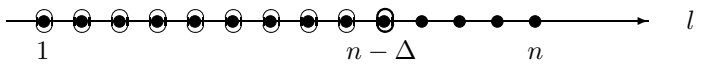

Fig. 8. Decoding error probability at $n - \Delta$ can be union bounded by the sum of probabilities of first decoding error at l, $1 \leq l \leq n - \Delta$. The dominant error event $p_n(n - \Delta)$ is the one in the highlighted oval(shortest delay).

Proof:

$$
\begin{aligned}
p_n(l) &= \sum_{x^n} \Pr\left[\exists\, \tilde{x}^n \in \mathcal{B}_x(x^n) \cap \mathcal{F}_n(l, x^n) \text{ s.t. } p_{\mathbf{x}}(\tilde{x}^n) \geq p_{\mathbf{x}}(x^n)\right] p_{\mathbf{x}}(x^n) \\
&\leq \sum_{x^n} \min\left[1, \sum_{\substack{\tilde{x}^n \in \mathcal{F}_n(l,\, x^n) \text{s.t.} \\ p_{\mathbf{x}}(x^n) \leq p_{\mathbf{x}}(\tilde{x}^n)}} \Pr[\tilde{x}^n \in \mathcal{B}_x(x^n)]\right] p_{\mathbf{x}}(x^n) && (30) \\
&= \sum_{x^{l-1}, x_l^n} \min\left[1, \sum_{\substack{\tilde{x}_l^n \text{ s.t.} \\ p_x(x_l^n) < p_x(\tilde{x}_l^n)}} \exp\{-(n-l+1)R_x\}\right] p_{\mathbf{x}}(x^{l-1}) p_{\mathbf{x}}(x_l^n) && (31) \\
&= \sum_{x_l^n} \min\left[1, \sum_{\substack{\tilde{x}_l^n \text{ s.t.} \\ p_x(x_l^n) < p_x(\tilde{x}_l^n)}} \exp\{-(n-l+1)R_x\}\right] p_{\mathbf{x}}(x_l^n) \\
&= \sum_{x_l^n} \min\left[1, \sum_{\tilde{x}_l^n} 1[p_x(\tilde{x}_l^n) > p_{\mathbf{x}}(x_l^n)] \exp\{-(n-l+1)R_x\}\right] p_{\mathbf{x}}(x_l^n) && (32) \\
&\leq \sum_{x_l^n} \min\left[1, \sum_{\tilde{x}_l^n} \min\left[1, \frac{p_x(\tilde{x}_l^n)}{p_x(x_l^n)}\right] \exp\{-(n-l+1)R_x\}\right] p_{\mathbf{x}}(x_l^n) \\
&\leq \sum_{x_l^n} \left[\sum_{\tilde{x}_l^n} \left[\frac{p_x(\tilde{x}_l^n)}{p_x(x_l^n)}\right]^{\frac{1}{1+\rho}} \exp\{-(n-l+1)R_x\}\right]^{\rho} p_{\mathbf{x}}(x_l^n) && (33) \\
&= \sum_{x_l^n} p_{\mathbf{x}}(x_l^n)^{\frac{1}{1+\rho}} \left[\sum_{\tilde{x}_l^n} [p_x(\tilde{x}_l^n)]^{\frac{1}{1+\rho}}\right]^{\rho} \exp\{-(n-l+1)\rho R_x\} \\
&= \left[\sum_x p_x(x)^{\frac{1}{1+\rho}}\right]^{(n-l+1)} \left[\sum_x p_x(x)^{\frac{1}{1+\rho}}\right]^{(n-l+1)\rho} \exp\{-(n-l+1)\rho R_x\} && (34) \\
&= \left[\sum_x p_x(x)^{\frac{1}{1+\rho}}\right]^{(n-l+1)(1+\rho)} \exp\{-(n-l+1)\rho R_x\} \\
&= \exp\left\{-(n-l+1)\left[\rho R_x - (1+\rho)\ln\left(\sum_x p_x(x)^{\frac{1}{1+\rho}}\right)\right]\right\}. && (35)
\end{aligned}
$$

In (30) the union bound is applied. In (31) we use the fact that after the first symbol in which two sequences differ, the remaining parity bits are independent, and the fact that only the likelihood of the differing suffixes matter. That is, if $x^{l-1} = \tilde{x}^{l-1}$, then $p_{\mathbf{x}}(x^n) < p_{\mathbf{x}}(\tilde{x}^n)$ if and only if $p_{\mathbf{x}}(x_l^n) < p_{\mathbf{x}}(\tilde{x}_l^n)$. In (32) $1(\cdot)$ is the indicator function, taking the value one if the argument is true, and zero if it is false. We get (33) by limiting ρ to the range $0 \leq \rho \leq 1$ since the arguments of the minimization are both positive and upper-bounded by one. We use the iid property of the source, exchanging sums and products to get (34). The bound in (35) is true for all ρ in the range $0 \leq \rho \leq 1$. Maximizing (35) over ρ gives $p_n(l) \leq \exp\{-(n-l+1)E_{ML}(R_x)\}$ where $E_{ML}(R_x)\}$ is defined in Theorem 2, in particular (9). ■

Using Lemma 2 in (27) gives

$$
\begin{aligned}
\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] &\leq \sum_{l=1}^{n-\Delta} \exp\{-(n-l+1)E_{ML}(R_x)\} && (36) \\
&= \sum_{l=1}^{n-\Delta} \exp\{-(n-l+1-\Delta)E_{ML}(R_x)\} \exp\{-\Delta E_{ML}(R_x)\} \\
&\leq K_0 \exp\{-\Delta E_{ML}(R_x)\} && (37)
\end{aligned}
$$

In (37) we pull out the exponent in Δ. The remaining summation is a sum over decaying exponentials, can thus can be bounded by some constant K_0. This proves Theorem 2.

B. Error events and sequential decoding

To better understand the dominant error event in the sum (36), consider constructing the ML estimate in a symbol-by-symbol sequential manner. The decoder starts by first identifying as candidates those sequences whose parities match the received bit stream up to time n. If the encoder observes the length-n sequence $\mathbf{x} = \mathbf{x}$, this is $\{\bar{\mathbf{x}} \text{ s.t. } \bar{\mathbf{x}} \in \mathcal{B}_x(\mathbf{x})\}$. The lth symbol of the estimate, $\hat{x}_l$, is defined as

$$\hat{x}_l = w_l \quad \text{where} \quad \mathbf{w} = \underset{\bar{\mathbf{x}} \in \mathcal{B}_x(\mathbf{x}) \text{ s.t. } \bar{x}^{l-1} = \hat{x}^{l-1}}{\arg\max} p_{x_l^n}(\bar{x}_l^n). \tag{38}$$

The estimate thus produced is the maximum likelihood estimate because the decision regarding which pair of sequences is more likely depends only on which one's suffix is more likely.

This is a decision-directed decoder. Semi-hard[13] estimate are made sequentially for each symbol. These estimates are then fixed, and taken as true when estimating subsequent symbols. Each such hard-decision is analogous to a classic block-coding Slepian-Wolf problem. This is because we only need to decide between sequences that start to differ in the symbol we are trying to estimate—previous symbols have been fixed, and subsequent symbols are not yet in question. Thus, all sequences that could lead to different estimates of symbol l are binned independently for the remainder of the block. This is why the error exponent we derive in (37) equals Gallager's block coding exponent [9]. Since the error exponent for each block-decoding problem is the same, the dominant error event is the hard-decision with the shortest block-length. This symbol is the last symbol we need to estimate. Its block-length equals the estimation delay Δ. We revisit this story in Section VII when we consider Slepian-Wolf coding. In that context the dominant error event has some features that do not arise in block coding.

C. Universal decoding

In this section we prove Theorem 3. We use the sequential decoder introduced in Section V-B, but with minimum-entropy, rather than maximum-likelihood, decoding. That is,

$$\hat{x}_l = w_l[l] \quad \text{where} \quad w^n[l] = \underset{\bar{x}^n \in \mathcal{B}_x(x^n) \text{ s.t. } \bar{x}^{l-1} = \hat{x}^{l-1}}{\arg\min} H(\bar{x}_l^n). \tag{39}$$

We term this a minimum suffix-entropy decoder. The reason for using this decoder instead of the standard minimum block-entropy decoder is that the block-entropy decoder has a polynomial term in n (resulting from summing over the type classes) that multiplies the exponential decay in Δ. For n large, this polynomial can dominate. Using the minimum suffix-entropy decoder results in a polynomial term in Δ.

With this decoder, errors can only occur if there is some sequence $\tilde{x}^n$ such that (i) $\tilde{x}^n \in \mathcal{B}_x(x^n)$, (ii) $\tilde{x}^{l-1} = x^{l-1}$, and $\tilde{x}_l \neq x_l$, for some $l \leq n - \Delta$, and (iii) the empirical suffix entropy of $\tilde{x}_l^n$ is such that $H(\tilde{x}_l^n) < H(x_l^n)$. Building on the common core of the achievability (25)–(27) with the substitution of universal decoding in the place of maximum likelihood results in the following definition of $p_n(l)$ (cf. (40) with (29),

$$p_n(l) = \sum_{x^n} \Pr\left[\exists\, \tilde{x}^n \in \mathcal{B}_x(x^n) \cap \mathcal{F}_n(l, x^n) \text{ s.t. } H(\tilde{x}_l^n) \leq H(x_l^n)\right] p_{\mathbf{x}}(x^n) \tag{40}$$

The following lemma gives a bound on $p_n(l)$.

Lemma 3: For minimum suffix-entropy decoding, $p_n(l) \leq (n - l + 2)^{2|\mathcal{X}|} \exp\{-(n - l + 1) E_{UN}(R_x)\}$.

Proof: We define P^{n-l} to be the type of length-$(n-l+1)$ sequence x_l^n, and $\mathcal{T}_{P^{n-l}}$ to be the corresponding type class so that $x_l^n \in \mathcal{T}_{P^{n-l}}$. Analogous definitions hold for $\tilde{P}^{n-l}$ and $\tilde{x}_l^n$. We rewrite the constraint $H(\tilde{x}_l^n) < H(\tilde{x}_l^n)$

[13]Decisions are only "hard" for computational time. As soon as the next set of parities arrive and real-time advances, all the computations are done again.

as $H(\tilde{P}^{n-l}) < H(P^{n-l})$. Thus,

$$\begin{aligned}
p_n(l) =& \sum_{x^n} \Pr\left[\exists\, \tilde{x}^n \in \mathcal{B}_x(x^n) \cap \mathcal{F}_n(l, x^n) \text{ s.t. } H(\tilde{x}_l^n) \le H(x_l^n)\right] p_{\mathbf{x}}(x^n) \\
\le& \sum_{x_1^n} \min\Big[1, \sum_{\substack{\tilde{x}_1^n \in \mathcal{F}_n(l,\, x^n) \text{ s.t.} \\ H(\tilde{x}_l^n) \le H(x_l^n)}} \Pr[\tilde{x}_1^n \in \mathcal{B}_x(x_1^n)]\Big] p_{\mathbf{x}}(x^n) \\
=& \sum_{x_1^{l-1}, x_l^n} \min\Big[1, \sum_{\substack{\tilde{x}_l^n \text{ s.t.} \\ H(\tilde{x}_l^n) \le H(x_l^n)}} \exp\{-(n-l+1)R_x\}\Big] p_{\mathbf{x}}(x^{l-1}) p_{\mathbf{x}}(x_l^n) \\
=& \sum_{x_l^n} \min\Big[1, \sum_{\substack{\tilde{x}_l^n \text{ s.t.} \\ H(\tilde{x}_l^n) \le H(x_l^n)}} \exp\{-(n-l+1)R_x\}\Big] p_{\mathbf{x}}(x_l^n) \qquad (41) \\
=& \sum_{P^{n-l}} \sum_{x_l^n \in \mathcal{T}_{P^{n-l}}} \min\Big[1, \sum_{\substack{\tilde{P}^{n-l} \text{ s.t.} \\ H(\tilde{P}^{n-l}) \le H(P^{n-l})}} \sum_{\tilde{x}_l^n \in \mathcal{T}_{\tilde{P}^{n-l}}} \exp\{-(n-l+1)R_x\}\Big] p_{\mathbf{x}}(x_l^n) \qquad (42) \\
\le& \sum_{P^{n-l}} \sum_{x_{l+1}^n \in \mathcal{T}_{P^{n-l}}} \min\Big[1, (n-l+2)^{|\mathcal{X}|} \exp\{-(n-l)[R_x - H(P^{n-l})]\}\Big] p_{\mathbf{x}}(x_l^n) \qquad (43) \\
\le& (n-l+2)^{|\mathcal{X}|} \sum_{P^{n-l}} \sum_{x_l^n \in \mathcal{T}_{P^{n-l}}} \exp\{-(n-l+1)[|R_x - H(P^{n-l})|^+]\} \\
& \qquad \exp\{-(n-l+1)[D(P^{n-l}\|p_x) + H(P^{n-l})]\} \qquad (44) \\
\le& (n-l+2)^{|\mathcal{X}|} \sum_{P^{n-l}} \exp\{-(n-l+1) \inf_q [D(q\|p_x) + |R_x - H(q)|^+]\} \qquad (45) \\
\le& (n-l+2)^{2|\mathcal{X}|} \exp\{-(n-l+1)E_{UN}(R_x)\} \qquad (46)
\end{aligned}$$

In going from (42) to (43) first note that the argument of the inner-most summation (over $\tilde{x}_l^n$) does not depend on $\mathbf{x}$. We then use the following relations: (i) $\sum_{\tilde{x}_l^n \in \mathcal{T}_{\tilde{P}^{n-l}}} = |\mathcal{T}_{\tilde{P}^{n-l}}| \le \exp\{(n-l+1)H(\tilde{P}^{n-l})\}$, which is a standard bound on the size of the type class, (ii) $H(\tilde{P}^{n-l}) \le H(P^{n-l})$ by the minimum-suffix-entropy decoding rule, and (iii) the polynomial bound on the number of types, $|\{\tilde{P}^{n-l}\}| \le (n-l+2)^{|\mathcal{X}|}$. In (44) we recall the function definition $|\cdot|^+ \triangleq \max\{0, \cdot\}$. We pull the polynomial term out of the minimization and use $p_{\mathbf{x}}(x_l^n) = \exp\{-(n-l+1)[D(P^{n-l}\|p_x) + H(P^{n-l})]\}$ for all $p_{\mathbf{x}}(x_l^n) \in \mathcal{T}_{P^{n-l}}$. It is also in (44) that we see why we use a minimum suffix-entropy decoding rule instead of a minimum entropy decoding rule. If we had not marginalized out over x^{l-1} in (41) then we would have a polynomial term out front in terms of n rather than $n-l$, which for large n could dominate the exponential decay in $n-l$. As the expression in (45) no longer depends on x_l^n, we simplify by using $|\mathcal{T}_{P^{n-l}}| \le \exp\{(n-l+1)H(P^{n-l})\}$. In (46) we use the definition of the universal error exponent $E_{UN}(R_x)$ from (10) of Theorem 3, and the polynomial bound on the number of types. ■

Lemma 3 and $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \le \sum_{l=1}^{n-\Delta} p_n(l)$ imply that:

$$\begin{aligned}
\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \le& \sum_{l=1}^{n-\Delta} (n-l+2)^{2|\mathcal{X}|} \exp\{-(n-l+1)E_{UN}(R_x)\} \\
\le& \sum_{l=1}^{n-\Delta} K_1 \exp\{-(n-l+1)[E_{UN}(R_x) - \gamma]\} \qquad (47) \\
\le& K_2 \exp\{-\Delta[E_{UN}(R_x) - \gamma]\} \qquad (48)
\end{aligned}$$

In (47) we incorporate the polynomial into the exponent. Namely, for all $a > 0$, $b > 0$, there exists a C such that $z^a \le C \exp\{b(z-1)\}$ for all $z \ge 1$.

We then make explicit the delay-dependent term. Pulling out the exponent in Δ, the remaining summation is a sum over decaying exponentials, and can be bounded by a constant. Together with K_1, this gives the constant K_2 in (48). This proves Theorem 3. Note that the γ in (48) does not enter the optimization because $\gamma > 0$ can be picked equal to any constant. The choice of γ effects the constant K in Theorem 3.

VI. Streaming source coding with side information at the decoder

If a random sequence y^n, related to the source x^n through a discrete memoryless channel, is observed at the decoder, then this side information can be used to reduce the rate of the source code. In this model $p_{\mathbf{x},\mathbf{y}}(x^n, y^n) = \prod_{i=1}^n p_{xy}(x_i, y_i) = \prod_{i=1}^n p_{x|y}(x_i|y_i)p_y(y_i)$. The source x^n is observed at the encoder, and the decoder, which observes y^n and a bit stream from the encoder, wants to estimate each source symbol x_i with a probability of error that decreases exponentially in the decoding delay Δ.

We can apply the analysis of Section V to this problem with a few minor modifications. For ML decoding, we need to pick the sequence with the maximum conditional probability given y^n. The error exponent can be derived using a similar Chernoff bounding argument as in section V. For universal decoding, the only change is that we now use a minimum suffix conditional-entropy decoder that compares sequence pairs $(\bar{x}^n, y^n)$ and $(\bar{\bar{x}}^n, y^n)$. In terms of the analysis, one change enters in (25) where we must also sum over the possible side information sequences. And in (42) the entropy condition in the summation over $\tilde{\mathbf{x}}$ changes to $H(\tilde{x}_{l+1}^n|y_{l+1}^n) < H(x_{l+1}^n|y_{l+1}^n)$ (or the equivalent type notation). Since there is no ambiguity in the side information, since y^n is observed at the decoder, this condition is equivalent to $H(\tilde{x}_{l+1}^n, y_{l+1}^n) < H(x_{l+1}^n, y_{l+1}^n)$.

These results are summarized in Theorems 4 and 5. We do not include the full derivation of these theorems as no new ideas are required.

VII. Streaming Slepian-Wolf source coding

In this section we provide the proofs of Theorems 6 and 7, which consider the two-user[14] Slepian-Wolf problem. As with the proofs of Theorems 2 and 3 in Sections V-A and V-C, we start by developing the common core of the proof in the context of maximum likelihood decoding. This allows us to develop the results for universal decoding more quickly and transparently. Furthermore, as shown in Theorem 8, maximum likelihood decoding and universal decoding provide the same reliability with delay.

A. Maximum Likelihood Decoding

In Theorems 6 and 7 three error events are considered: (i) $\Pr[x^{n-\Delta} \neq \hat{x}^{n-\Delta}]$, (ii) $\Pr[y^{n-\Delta} \neq \hat{y}^{n-\Delta}]$, and (iii) $\Pr[(x^{n-\Delta}, y^{n-\Delta}) \neq (\hat{x}^{n-\Delta}, \hat{y}^{n-\Delta})]$. We develop the error exponent for case (i). The error exponent for case (ii) follows from a similar derivation, and that of case (iii) from an application of the union bound resulting in an exponent that is the minimum of the exponents of cases (i) and (ii).

To lead to the decoding error $\Pr[x^{n-\Delta} \neq \hat{x}^{n-\Delta}]$ there must be some spurious source pair $(\tilde{x}^n, \tilde{y}^n)$ that satisfies three conditions: (i) $\tilde{x}^n \in \mathcal{B}_x(x^n)$ and $\tilde{y}^n \in \mathcal{B}_y(y^n)$, (ii) it must be more likely than the true pair $p_{\mathbf{x},\mathbf{y}}(\tilde{x}^n, \tilde{y}^n) > p_{\mathbf{x},\mathbf{y}}(x^n, y^n)$, and (iii) $\tilde{x}_l \neq x_l$ for some $l \leq n - \Delta$.

The error probability is

$$\begin{aligned}
\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] &= \sum_{x^n, y^n} \Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta} | x^n = x^n, y^n = y^n] p_{\mathbf{x},\mathbf{y}}(x^n, y^n) \\
&\leq \sum_{x^n, y^n} p_{\mathbf{x},\mathbf{y}}(x^n, y^n) \Big\{ \sum_{l=1}^{n-\Delta} \sum_{k=1}^{n+1} \\
&\qquad \Pr\left[\exists\, (\tilde{x}^n, \tilde{y}^n) \in \mathcal{B}_x(x^n) \times \mathcal{B}_y(y^n) \cap \mathcal{F}_n(l, k, x^n, y^n) \text{ s.t. } p_{\mathbf{x},\mathbf{y}}(\tilde{x}^n, \tilde{y}^n) \geq p_{\mathbf{x},\mathbf{y}}(x^n, y^n)\right] \Big\} \qquad (49) \\
&= \sum_{l=1}^{n-\Delta} \sum_{k=1}^{n+1} \Big\{ \sum_{x^n, y^n} p_{\mathbf{x},\mathbf{y}}(x^n, y^n) \\
&\qquad \Pr\left[\exists\, (\tilde{x}^n, \tilde{y}^n) \in \mathcal{B}_x(x^n) \times \mathcal{B}_y(y^n) \cap \mathcal{F}_n(l, k, x^n, y^n) \text{ s.t. } p_{\mathbf{x},\mathbf{y}}(\tilde{x}^n, \tilde{y}^n) \geq p_{\mathbf{x},\mathbf{y}}(x^n, y^n)\right] \Big\} \\
&= \sum_{l=1}^{n-\Delta} \sum_{k=1}^{n+1} p_n(l, k). \qquad (50)
\end{aligned}$$

[14]The multiuser case is essentially the same, just with a lot more notation and minimization parameters $\gamma_1, \gamma_2, \ldots$.

In (49) we decompose the error event into a number of mutually exclusive events by partitioning all source pairs $(\tilde{x}^n, \tilde{y}^n)$ into sets $\mathcal{F}_n(l,k,x^n,y^n)$ defined by the times l and k at which $\tilde{x}^n$ and $\tilde{y}^n$ diverge from the realized source sequences. The set $\mathcal{F}_n(l,k,x^n,y^n)$ is defined as

$$\mathcal{F}_n(l,k,x^n,y^n) = \{(\bar{x}^n,\tilde{y}^n) \in \mathcal{X}^n \times \mathcal{Y}^n \text{ s.t. } \bar{x}^{l-1} = x^{l-1}, \bar{x}_l \neq x_l, \bar{y}^{k-1} = y^{k-1}, \bar{y}_k \neq y_k\}, \tag{51}$$

In contrast to streaming point-to-point or side-information coding (cf. (51) with (28)), the partition is now doubly-indexed. To find the dominant error event, we must search over both indices. Having two dimensions to search over results in an extra minimization when calculating the error exponent (and leads to the infimum over γ in Theorem 6).

Finally, to get (50) we define $p_n(l,k)$ as

$$\begin{aligned} & p_n(l,k) \\ = & \sum_{x^n,y^n} p_{\mathbf{x},\mathbf{y}}(x^n,y^n) \Pr\Big[\exists\, (\tilde{x}^n,\tilde{y}^n) \in \mathcal{B}_x(x^n) \times \mathcal{B}_y(y^n) \cap \mathcal{F}_n(l,k,x^n,y^n) \text{ s.t. } p_{\mathbf{x},\mathbf{y}}(\tilde{x}^n,\tilde{y}^n) \geq p_{\mathbf{x},\mathbf{y}}(x^n,y^n)\Big]. \end{aligned}$$

The following lemma provides an upper bound on $p_n(l,k)$:

Lemma 4:

$$\begin{aligned} p_n(l,k) &\leq \exp\{-(n-l+1)E_x(R_x,R_y,\tfrac{k-l}{n-l+1})\} \quad \text{if} \quad l \leq k, \\ p_n(l,k) &\leq \exp\{-(n-k+1)E_y(R_x,R_y,\tfrac{l-k}{n-k+1})\} \quad \text{if} \quad l \geq k, \end{aligned} \tag{52}$$

where $E_x(R_x,R_y,\gamma)$ and $E_y(R_x,R_y,\gamma)$ are defined in (13) and (14) respectively. Notice that $l,k \leq n$, for $l \leq k$: $\frac{k-l}{n-l+1} \in [0,1]$ serves as γ in the error exponent $E_x(R_x,R_y,\gamma)$. Similarly for $l \geq k$.

Proof: The bound depends on whether $l \leq k$ or $l \geq k$. Consider the case for $l \leq k$,

$$\begin{aligned} & p_n(l,k) \\ & = \sum_{x^n,y^n} p_{\mathbf{x},\mathbf{y}}(x^n,y^n) \Pr[\exists\, (\tilde{x}^n,\tilde{y}^n) \in \mathcal{B}_x(x^n) \times \mathcal{B}_y(y^n) \cap \mathcal{F}_n(l,k,x^n,y^n) \text{ s.t. } p_{\mathbf{x},\mathbf{y}}(x^n,y^n) < p_{\mathbf{x},\mathbf{y}}(\tilde{x}^n,\tilde{y}^n)] \\ & \leq \sum_{x^n,y^n} \min\Big[1, \sum_{\substack{(\tilde{x}^n,\tilde{y}^n) \in \mathcal{F}_n(l,k,x^n,y^n) \\ p_{\mathbf{x},\mathbf{y}}(x^n,y^n) < p_{\mathbf{x},\mathbf{y}}(\tilde{x}^n,\tilde{y}^n)}} \Pr[\tilde{x}^n \in \mathcal{B}_x(x^n), \tilde{y}^n \in \mathcal{B}_y(y^n)]\Big] p_{\mathbf{x},\mathbf{y}}(x^n,y^n) \qquad (53) \\ & \leq \sum_{x_l^n,y_l^n} \min\Big[1, \sum_{\substack{(\tilde{x}_l^n,\tilde{y}_l^n) \text{ s.t. } \tilde{y}^{k-1} = y^{k-1} \\ p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n) < p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^n,\tilde{y}_l^n)}} \exp\{-(n-l+1)R_x - (n-k+1)R_y\}\Big] p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n) \qquad (54) \\ & = \sum_{x_l^n,y_l^n} \min\Big[1, \sum_{\tilde{x}_l^n,\tilde{y}_k^n} \exp\{-(n-l+1)R_x - (n-k+1)R_y\} \\ & \qquad\qquad \mathbb{1}[p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^{k-1},y_l^{k-1}) p_{\mathbf{x},\mathbf{y}}(\tilde{x}_k^n,\tilde{y}_k^n) > p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n)]\Big] p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n) \\ & \leq \sum_{x_l^n,y_l^n} \min\Bigg[1, \sum_{\tilde{x}_l^n,\tilde{y}_k^n} \exp\{-(n-l+1)R_x - (n-k+1)R_y\} \\ & \qquad\qquad \min\left[1, \frac{p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^{k-1},y_l^{k-1}) p_{\mathbf{x},\mathbf{y}}(\tilde{x}_k^n,\tilde{y}_k^n)}{p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n)}\right]\Bigg] p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n) \\ & \leq \sum_{x_l^n,y_l^n} \left[\sum_{\tilde{x}_l^n,\tilde{y}_k^n} e^{-(n-l+1)R_x-(n-k+1)R_y} \left[\frac{p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^{k-1},y_l^{k-1}) p_{\mathbf{x},\mathbf{y}}(\tilde{x}_k^n,\tilde{y}_k^n)}{p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n)}\right]^{\frac{1}{1+\rho}}\right]^{\rho} p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n) \qquad (55) \\ & = e^{-(n-l+1)\rho R_x-(n-k+1)\rho R_y} \sum_{x_l^n,y_l^n} \left[\sum_{\tilde{x}_l^n,\tilde{y}_k^n} [p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^{k-1},y_l^{k-1}) p_{\mathbf{x},\mathbf{y}}(\tilde{x}_k^n,\tilde{y}_k^n)]^{\frac{1}{1+\rho}}\right]^{\rho} p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n)^{\frac{1}{1+\rho}} \end{aligned}$$

$$= e^{-(n-l+1)\rho R_x-(n-k+1)\rho R_y} \sum_{y_l^{k-1}} \Big[\sum_{x_l^{k-1}} p_{\mathbf{x},\mathbf{y}}(x_l^{k-1}, y_l^{k-1})^{\frac{1}{1+\rho}} \Big] \Big[\sum_{\tilde{x}_l^{k-1}} p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^{k-1}, y_l^{k-1})^{\frac{1}{1+\rho}} \Big]^{\rho}$$
$$\Big[\sum_{\tilde{x}_k^n, \tilde{y}_k^n} p_{\mathbf{x},\mathbf{y}}(\tilde{x}_k^n, \tilde{y}_k^n)^{\frac{1}{1+\rho}} \Big]^{\rho} \sum_{x_k^n, y_k^n} p_{\mathbf{x},\mathbf{y}}(x_k^n, y_k^n)^{\frac{1}{1+\rho}}$$

$$= e^{-(n-l+1)\rho R_x-(n-k+1)\rho R_y} \left[\sum_{y_l^{k-1}} \Big[\sum_{x_l^{k-1}} p_{\mathbf{x},\mathbf{y}}(x_l^{k-1}, y_l^{k-1})^{\frac{1}{1+\rho}} \Big]^{1+\rho} \right] \Big[\sum_{x_k^n, y_k^n} p_{\mathbf{x},\mathbf{y}}(x_k^n, y_k^n)^{\frac{1}{1+\rho}} \Big]^{1+\rho}$$

$$= e^{-(n-l+1)\rho R_x-(n-k+1)\rho R_y} \left[\sum_{y} \Big[\sum_{x} p_{\mathsf{x},\mathsf{y}}(x, y)^{\frac{1}{1+\rho}} \Big]^{1+\rho} \right]^{k-l} \Big[\sum_{x,y} p_{\mathsf{x},\mathsf{y}}(x, y)^{\frac{1}{1+\rho}} \Big]^{(1+\rho)(n-k+1)} \tag{56}$$

$$= \exp\left\{ -(k-l) \left[\rho R_x - \log \Big[\sum_{y} \Big[\sum_{x} p_{\mathsf{x},\mathsf{y}}(x, y)^{\frac{1}{1+\rho}} \Big]^{1+\rho} \Big] \right] \right\}$$
$$\exp\left\{ -(n-k+1) \left[\rho(R_x + R_y) - (1+\rho) \log \Big[\sum_{x,y} p_{\mathsf{x},\mathsf{y}}(x, y)^{\frac{1}{1+\rho}} \Big] \right] \right\}$$

$$= \exp\big\{ -(k-l) E_{x|y}(R_x, \rho) - (n-k+1) E_{xy}(R_x, R_y, \rho) \big\} \tag{57}$$

$$= \exp\left\{ -(n-l+1) \Big[\frac{k-l}{n-l+1} E_{x|y}(R_x, \rho) + \frac{n-k+1}{n-l+1} E_{xy}(R_x, R_y, \rho) \Big] \right\} \tag{58}$$

$$\leq \exp\left\{ -(n-l+1) \sup_{\rho \in [0,1]} \Big[\frac{k-l}{n-l+1} E_{x|y}(R_x, \rho) + \frac{n-k+1}{n-l+1} E_{xy}(R_x, R_y, \rho) \Big] \right\} \tag{59}$$

$$= \exp\left\{ -(n-l+1) E_x^{ML}\left(R_x, R_y, \frac{k-l}{n-l+1} \right) \right\} = \exp\left\{ -(n-l+1) E_x(R_x, R_y, \frac{k-l}{n-l+1}) \right\}. \tag{60}$$

In (53) we explicitly indicate the three conditions that a suffix pair $(\tilde{x}_l^n, \tilde{y}_k^n)$ must satisfy to result in a decoding error. In (54) we sum out over the common prefixes (x^{l-1}, y^{l-1}), and use the fact that the random binning is done independently at each encoder, see Definition. 2. We get (55) by limiting ρ to the interval $0 \leq \rho \leq 1$, as in (33). Getting (56) from (55) follows by a number of basic manipulations. In (56) we get the single letter expression by again using the memoryless property of the sources. In (57) we use the definitions of $E_{x|y}$ and E_{xy} from (14) of Theorem 6. Noting that the bound holds for all $\rho \in [0, 1]$ optimizing over ρ results in (59). Finally, using the definition of (13) and the remark following Theorem 8 that the maximum-likelihood and universal exponents are equal gives (60). The bound on $p_n(l, k)$ when $l > k$, is developed in an analogous fashion. ■

We use Lemma 4 together with (50) to bound $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}]$ for two distinct cases. The first, simpler case, is when $\inf_{\gamma \in [0,1]} E_y(R_x, R_y, \gamma) > \inf_{\gamma \in [0,1]} E_x(R_x, R_y, \gamma)$. To bound $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}]$ in this case, we split the sum over the $p_n(l, k)$ into two terms, as visualized in Fig 9. There are $(n+1) \times (n-\Delta)$ such events to account for (those inside the box). The probability of the event within each oval are summed together to give an upper bound on $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}]$. We add extra probabilities outside of the box but within the ovals to make the summation symmetric thus simpler. Those extra error events do not impact the error exponent because $\inf_{\gamma \in [0,1]} E_y(R_x, R_y, \rho, \gamma) \geq \inf_{\gamma \in [0,1]} E_x(R_x, R_y, \rho, \gamma)$. The possible dominant error events are highlighted in

Figure 9 . Thus,

$$\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \leq \sum_{l=1}^{n-\Delta}\sum_{k=l}^{n+1} p_n(l,k) + \sum_{k=1}^{n-\Delta}\sum_{l=k}^{n+1} p_n(l,k) \tag{61}$$

$$\leq \sum_{l=1}^{n-\Delta}\sum_{k=l}^{n+1} \exp\{-(n-l+1)\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)\} + \sum_{k=1}^{n-\Delta}\sum_{l=k}^{n+1} \exp\{-(n-k+1)\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)\} \tag{62}$$

$$= \sum_{l=1}^{n-\Delta}\Big[(n-l+2)\exp\{-(n-l+1)\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)\}$$

$$+ \sum_{k=1}^{n-\Delta}\Big[(n-k+2)\exp\{-(n-k+1)\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)\}$$

$$\leq 2\sum_{l=1}^{n-\Delta}\Big[(n-l+2)\exp\{-(n-l+1)\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)\} \tag{63}$$

$$\leq \sum_{l=1}^{n-\Delta} C_1\exp\{-(n-l+2)[\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)-\alpha]\} \tag{64}$$

$$\leq C_2\exp\{-\Delta[\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)-\alpha]\} \tag{65}$$

Equation (61) follows directly from (50), in the first term $l \leq k$, in the second term $l \geq k$. In (62), we use Lemma 4. In (63) we use the assumption that $\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma) > \inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)$. In (64) the $\alpha > 0$ results from incorporating the polynomial into the first exponent, and can be chosen as small as desired. Combining terms and summing out the decaying exponential yield the bound (65).

The second, more involved case, is when $\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\rho,\gamma) < \inf_{\gamma\in[0,1]} E_x(R_x,R_y,\rho,\gamma)$. To bound $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}]$, we could use the same bounding technique used in the first case. This gives the error exponent $\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)$ which is generally smaller than what we can get by dividing the error events in a new scheme as shown in Figure 10. In this situation we split (50) into three terms, as visualized in Fig 10. Just as in the first case shown in Fig 9, there are $(n+1)\times(n-\Delta)$ such events to account for (those inside the box). The error events are partitioned into 3 regions. Region 2 and 3 are separated by $k^*(l)$ using a dotted line. In region 3, we add extra probabilities outside of the box but within the ovals to make the summation simpler. Those extra error events do not affect the error exponent as shown in the proof. The possible dominant error events are highlighted shown in Fig 10. Thus,

$$\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \leq \sum_{l=1}^{n-\Delta}\sum_{k=l}^{n+1} p_n(l,k) + \sum_{l=1}^{n-\Delta}\sum_{k=k^*(l)}^{l-1} p_n(l,k) + \sum_{l=1}^{n-\Delta}\sum_{k=1}^{k^*(l)-1} p_n(l,k) \tag{66}$$

Where $\sum_{k=1}^{0} p_k = 0$. The lower boundary of Region 2 is $k^*(l) \geq 1$ as a function of n and l:

$$k^*(l) = \max\left\{1, n+1-\lceil\frac{\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)}{\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)}\rceil(n+1-l)\right\} = \max\{1, n+1-G(n+1-l)\} \tag{67}$$

where we use G to denote the ceiling of the ratio of exponents. Note that when $\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma) > \inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)$ then $G = 1$ and region two of Fig. 10 disappears. In other words, the middle term of (66) equals zero. This is the first case considered. We now consider the cases when $G \geq 2$ (because of the ceiling function G is a positive integer).

The first term of (66), i.e., region one in Fig. 10 where $l \leq k$, is bounded in the same way that the first term of (61) is, giving

$$\sum_{l=1}^{n-\Delta}\sum_{k=l}^{n+1} p_n(l,k) \leq C_2\exp\{-\Delta[\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)-\alpha]\}. \tag{68}$$

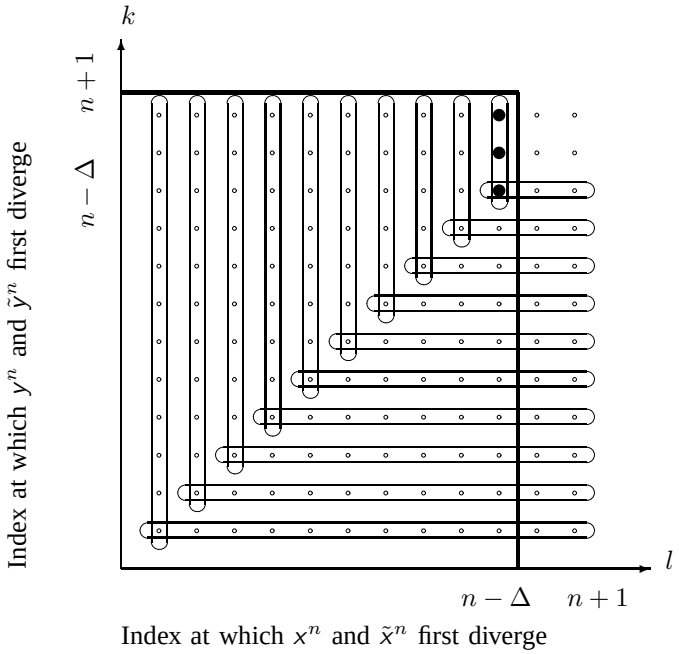

Fig. 9. Two dimensional plot of the error probabilities $p_n(l,k)$, corresponding to error events (l,k), contributing to $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}]$ in the situation where $\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\rho,\gamma) \geq \inf_{\gamma\in[0,1]} E_x(R_x,R_y,\rho,\gamma)$.

In Fig. 10, region two is upper bounded by the 45-degree line, and lower bounded by $k^*(l)$. The second term of (66), corresponding to this region where $l \geq k$,

$$\begin{aligned}
\sum_{l=1}^{n-\Delta}\sum_{k=k^*(l)}^{l-1} p_n(l,k) &\leq \sum_{l=1}^{n-\Delta}\sum_{k=k^*(l)}^{l-1} \exp\{-(n-k+1)E_y(R_x,R_y,\frac{l-k}{n-k+1})\} \\
&= \sum_{l=1}^{n-\Delta}\sum_{k=k^*(l)}^{l-1} \exp\{-(n-k+1)\frac{n-l+1}{n-l+1}E_y(R_x,R_y,\frac{l-k}{n-k+1})\} \qquad (69)\\
&\leq \sum_{l=1}^{n-\Delta}\sum_{k=k^*(l)}^{l-1} \exp\{-(n-l+1)\inf_{\gamma\in[0,1]}\frac{1}{1-\gamma}E_y(R_x,R_y,\gamma)\} \qquad (70)\\
&= \sum_{l=1}^{n-\Delta}(l-k^*(l))\exp\{-(n-l+1)\inf_{\gamma\in[0,1]}\frac{1}{1-\gamma}E_y(R_x,R_y,\gamma)\} \qquad (71)
\end{aligned}$$

In (69) we note that $l \geq k$, so define $\frac{l-k}{n-k+1} = \gamma$ as in (70). Then $\frac{n-k+1}{n-l+1} = \frac{1}{1-\gamma}$.

The third term of (66), i.e., the intersection of region three and the "box" in Fig. 10 where $l \geq k$, can be bounded

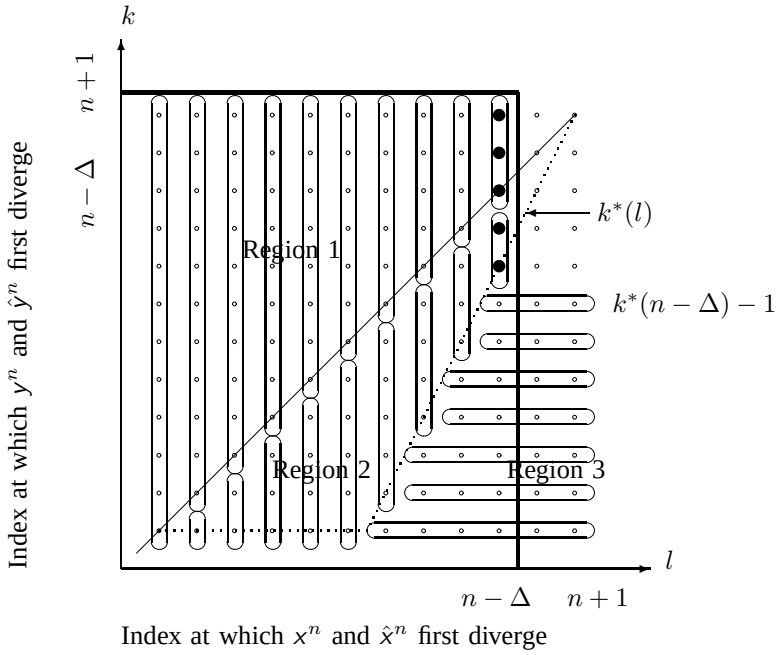

Fig. 10. Two dimensional plot of the error probabilities $p_n(l,k)$, corresponding to error events (l,k), contributing to $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}]$ in the situation where $\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma) < \inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)$.

as,

$$\sum_{l=1}^{n-\Delta}\sum_{k=1}^{k^*(l)-1} p_n(l,k) \leq \sum_{l=1}^{n+1}\sum_{k=1}^{\min\{l,k^*(n-\Delta)-1\}} p_n(l,k) \tag{72}$$

$$= \sum_{k=1}^{k^*(n-\Delta)-1}\sum_{l=k}^{n+1} p_n(l,k) \tag{73}$$

$$\leq \sum_{k=1}^{k^*(n-\Delta)-1}\sum_{l=k}^{n+1} \exp\{-(n-k+1)E_y(R_x,R_y,\frac{l-k}{n-k+1})\}$$

$$\leq \sum_{k=1}^{k^*(n-\Delta)-1}\sum_{l=k}^{n+1} \exp\{-(n-k+1)\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)\}$$

$$\leq \sum_{k=1}^{k^*(n-\Delta)-1} (n-k+2)\exp\{-(n-k+1)\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)\} \tag{74}$$

In (72) we note that $l \leq n-\Delta$ thus $k^*(n-\Delta)-1 \geq k^*(l)-1$, also $l \geq 1$, so $l \geq k^*(l)-1$. This can be visualized in Fig 10 as we extend the summation from the intersection of the "box" and region 3 to the whole region under the diagonal line and the horizontal line $k = k^*(n-\Delta)-1$. In (73) we simply switch the order of the summation.

Finally when $G \geq 2$, we substitute (68), (71), and (74) into (66) to give

$$\begin{aligned}
\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] &\leq C_2 \exp\{-\Delta[\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma) - \alpha]\} \\
&+ \sum_{l=1}^{n-\Delta} (l-k^*(l)) \exp\{-(n-l+1)\inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x,R_y,\gamma)\} \qquad (75)\\
&+ \sum_{k=1}^{k^*(n-\Delta)-1} (n-k+2)\exp\{-(n-k+1)\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)\} \\
&\leq C_2 \exp\{-\Delta[\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma) - \alpha]\} \\
&+ \sum_{l=1}^{n-\Delta} (l-n-1+G(n+1-l)) \exp\{-(n-l+1)\inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x,R_y,\gamma)\} \\
&+ \sum_{k=1}^{n+1-G(\Delta+1)} (n-k+2)\exp\{-(n-k+1)\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)\} \qquad (76)\\
&\leq C_2 \exp\{-\Delta[\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma) - \alpha]\} \\
&+ (G-1)C_3 \exp\{-\Delta\big[\inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x,R_y,\gamma) - \alpha\big]\} \\
&+ C_4 \exp\{-\big[\Delta G \inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma) - \alpha\big]\} \\
&\leq C_5 \exp\Big\{-\Delta\Big[\min\Big\{\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma), \inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x,R_y,\gamma)\Big\} - \alpha\Big]\Big\}. \qquad (77)
\end{aligned}$$

To get (76), we use the fact that $k^*(l) \geq n+1-G(n+1-l)$ from the definition of $k^*(l)$ in (67) to upper bound the second term. We exploit the definition of G to convert the exponent in the third term to $\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)$. Finally, to get (77) we gather the constants together, sum out over the decaying exponentials, and are limited by the smaller of the two exponents.

Note: in the proof of Theorem 6, we regularly double count the error events or add smaller extra probabilities to make the summations simpler. But it should be clear that the error exponent is not affected.

B. Universal Decoding

As discussed in Section V-C, we do not use a pairwise minimum joint-entropy decoder because of polynomial term in n would multiply the exponential decay in Δ. Analogous to the sequential decoder used there, we use a "weighted suffix entropy" decoder. The decoding starts by first identifying candidate sequence pairs as those that agree with the encoding bit streams up to time n, i.e., $\bar{x}^n \in \mathcal{B}_x(x^n), \bar{y}^n \in \mathcal{B}_y(y^n)$. For any one of the $|\mathcal{B}_x(x^n)||\mathcal{B}_y(y^n)|$ sequence pairs in the candidate set, i.e., $(\bar{x}^n,\bar{y}^n) \in \mathcal{B}_x(x^n)\times\mathcal{B}_y(y^n)$ we compute $(n+1)\times(n+1)$ weighted entropies:

$$\begin{aligned}
H_S(l,k,\bar{x}^n,\bar{y}^n) &= H(\bar{x}_l^{(n+1-l)},\bar{y}_l^{(n+1-l)}), \quad l=k \\
H_S(l,k,\bar{x}^n,\bar{y}^n) &= \frac{k-l}{n+1-l} H(\bar{x}_l^{k-1}|\bar{y}_l^{k-1}) + \frac{n+1-k}{n+1-l} H(\bar{x}_k^n,\bar{y}_k^n), \quad l<k \\
H_S(l,k,\bar{x}^n,\bar{y}^n) &= \frac{l-k}{n+1-k} H(\bar{y}_k^{l-1}|\bar{x}_k^{l-1}) + \frac{n+1-l}{n+1-k} H(\bar{x}_l^n,\bar{y}_l^n), \quad l>k.
\end{aligned}$$

We define the *score* of $(\bar{x}^n,\bar{y}^n)$ as the pair of integers $i_x(\bar{x}^n,\bar{y}^n)$, $i_y(\bar{x}^n,\bar{y}^n)$ s.t.,

$$\begin{aligned}
i_x(\bar{x}^n,\bar{y}^n) &= \max\{i : H_S(l,k,(\bar{x}^n,\bar{y}^n)) < H_S(l,k,\tilde{x}^n,\tilde{y}^n) \forall k=1,2,...n+1, \forall l=1,2,...i, \\
&\quad \forall(\tilde{x}^n,\tilde{y}^n) \in \mathcal{B}_x(x^n)\times\mathcal{B}_y(y^n) \cap \mathcal{F}_n(l,k,\bar{x}^n,\bar{y}^n)\} \qquad (78)\\
i_y(\bar{x}^n,\bar{y}^n) &= \max\{i : H_S(l,k,(\bar{x}^n,\bar{y}^n)) < H_S(l,k,\tilde{x}^n,\tilde{y}^n) \forall l=1,2,...n+1, \forall k=1,2,...i, \\
&\quad \forall(\tilde{x}^n,\tilde{y}^n) \in \mathcal{B}_x(x^n)\times\mathcal{B}_y(y^n) \cap \mathcal{F}_n(l,k,\bar{x}^n,\bar{y}^n)\} \qquad (79)
\end{aligned}$$

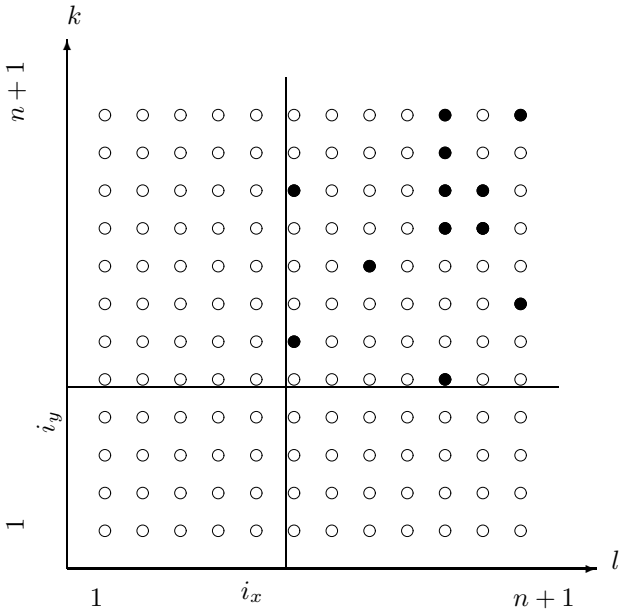

Fig. 11. 2D interpretation of the *score*, $(i_x(\bar{x}^n, \bar{y}^n), i_y(\bar{x}^n, \bar{y}^n))$, of a sequence pair $(\bar{x}^n, \bar{y}^n)$. If there exists a sequence pair in $\mathcal{F}_n(l, k, \bar{x}^n, \bar{y}^n)$ with less or the same score, then (l, k) is marked with a solid dot. The *score* $i_x(\bar{x}^n, \bar{y}^n)$ is the largest integer which is smaller than all the x-coordinates of the marked points. Similarly for $i_y(\bar{x}^n, \bar{y}^n)$,

While $\mathcal{F}_n(l, k, x^n, y^n)$ is the same set as defined in (51), we repeat the definition here for convenience,

$$\mathcal{F}_n(l, k, x^n, y^n) = \{(\bar{x}^n, \tilde{y}^n) \in \mathcal{X}^n \times \mathcal{Y}^n \text{ s.t. } \bar{x}^{l-1} = x^{l-1}, \bar{x}_l \neq x_l, \bar{y}^{k-1} = y^{k-1}, \bar{y}_k \neq y_k\}.$$

The definition of $(i_x(\bar{x}^n, \bar{y}^n), i_y(\bar{x}^n, \bar{y}^n))$ can be visualized in the following procedure. As shown in Fig. 11, for all $1 \leq l, k \leq n+1$, if there exists $(\bar{\bar{x}}^n, \bar{\bar{y}}^n) \in \mathcal{F}_n(l, k, (\bar{x}^n, \bar{y}^n)) \cap \mathcal{B}_x(x^n) \times \mathcal{B}_y(y^n)$ s.t. $H_S(l, k, \bar{x}^n, \bar{y}^n) \geq H_S(l, k, \bar{\bar{x}}^n, \bar{\bar{y}}^n)$, then we mark (l, k) on the plane as shown in Fig.11. Eventually we pick the maximum integer which is smaller than all marked x-coordinates as $i_x(\bar{x}^n, \bar{y}^n)$ and the maximum integer which is smaller than all marked y-coordinates as $i_y(\bar{x}^n, \bar{y}^n)$. The score of $(\bar{x}^n, \bar{y}^n)$ tells us the first branch(either x or y) point where a "better sequence pair" (with a smaller weighted entropy) exists.

Define the set of the winners as the sequences (not sequence pair) with the maximum score:

$$\mathcal{W}_n^x = \{\bar{x}^n \in \mathcal{B}_x(x^n) : \exists \bar{y}^n \in \mathcal{B}_y(y^n), s.t. i_x(\bar{x}^n, \bar{y}^n) \geq i_x(\tilde{x}^n, \tilde{y}^n), \forall (\tilde{x}^n, \tilde{y}^n) \in \mathcal{B}_x(x^n) \times \mathcal{B}_y(y^n)\}$$
$$\mathcal{W}_n^y = \{\bar{y}^n \in \mathcal{B}_y(y^n) : \exists \bar{x}^n \in \mathcal{B}_x(x^n), s.t. i_y(\bar{x}^n, \bar{y}^n) \geq i_y(\tilde{x}^n, \tilde{y}^n), \forall (\tilde{x}^n, \tilde{y}^n) \in \mathcal{B}_x(x^n) \times \mathcal{B}_y(y^n)\}$$

Then arbitrarily pick one sequence from $\mathcal{W}_n^x$ and one from $\mathcal{W}_n^y$ as the decision $(\hat{x}^n, \hat{y}^n)$.

We bound the probability that there exists a sequence pair in $\mathcal{F}_n(l, k, (x^n, y^n)) \cap \mathcal{B}_x(x^n) \times \mathcal{B}_y(y^n)$ with smaller weighted minimum-entropy suffix score as:

$$\begin{aligned} p_n(l, k) &= \sum_{x^n} \sum_{y^n} p_{xy}(x^n, y^n) P(\exists (\tilde{x}_1^n, \tilde{y}_1^n) \in \mathcal{B}_x(x^n) \times \mathcal{B}_y(y^n) \cap \mathcal{F}_n(l, k, x^n, y^n), \\ &\quad s.t. H_S(l, k, \tilde{x}^n, \tilde{y}^n) \leq H_S(l, k, (x^n, y^n))) \end{aligned}$$

Note that the $p_n(l, k)$ here differs from the $p_n(l, k)$ defined in the ML decoding by replacing $p_{xy}(x^n, y^n) \leq p_{xy}(\tilde{x}^n, \tilde{y}^n)$ with $H_S(l, k, \tilde{x}^n, \tilde{y}^n) \leq H_S(l, k, (x^n, y^n))$.

The following lemma, analogous to (50) for ML decoding, tells us that the "suffix weighted entropy" decoding rule is a good one.

Lemma 5: Upper bound on symbol-wise decoding error $P_{ex}(k, k+d)$:

$$\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \leq \sum_{l=1}^{n-\Delta} \sum_{k=1}^{n+1} p_n(l,k)$$

Proof: According to the decoding rule, $\hat{x}^{n-\Delta} \neq x^{n-\Delta}$ implies that there exists a sequence $\tilde{x}^n \in \mathcal{W}_n^x$ s.t.$\tilde{x}^{n-\Delta} \neq x^{n-\Delta}$. This means that there exists a sequence $\tilde{y}^n \in \mathcal{B}_y(y^n)$, s.t. $i_x(\tilde{x}^n, \tilde{y}^n) \geq i_x(x^n, y^n)$. Suppose that $(\tilde{x}^n, \tilde{y}^n) \in \mathcal{F}_n(l,k,x^n,y^n)$, then $l \leq n-\Delta$ because $\tilde{x}^{n-\Delta} \neq x^{n-\Delta}$. By the definition of i_x, we know that $H_S(l,k,\tilde{x}^n,\tilde{y}^n) \leq H_S(l,k,x^n,y^n)$. And using the union bound argument we get the desired inequality. ■

We only need to bound each single error probability $p_n(l,k)$ to finish the proof.

Lemma 6: Upper bound on $p_n(l,k)$, $l \leq k$: $\forall \gamma > 0$, $\exists K_1 < \infty$, s.t.

$$p_n(l,k) \leq \exp\{-(n-l+1)[E_x(R_x,R_y,\lambda) - \gamma]\}$$

where $\lambda = (k-l)/(n-l+1) \in [0,1]$.

Proof: Here the error probability $p_n(l,k)$ can be thought as starting from (54) with the condition $(k-l)H(\tilde{x}_l^{k-1}|\tilde{y}_l^{k-1}) + (n-k+1)H(\tilde{x}_k^n,\tilde{y}_k^n) < (k-l)H(x_l^{k-1}|y_l^{k-1}) + (n-k+1)H(x_k^n,y_k^n)$ substituted for $p(\tilde{x}_l^n,\tilde{y}_l^n) > p(x_l^n,y_l^n)$, we get

$$p_n(l,k) = \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}} \sum_{\substack{y_l^{k-1} \in \mathcal{T}_{P^{k-l}}, \\ y_k^n \in \mathcal{T}_{P^{n-k}}}} \sum_{\substack{x_l^{k-1} \in \mathcal{T}_{V^{k-l}}(y_l^{k-1}), \\ x_k^n \in \mathcal{T}_{V^{n-k}}(y_k^n)}} \min\Big[1, \sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k} \text{ s.t.} \\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l}) < \\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} \sum_{\tilde{y}_k^n \in \mathcal{T}_{\tilde{P}^{n-k}}} \sum_{\tilde{x}_l^{k-1} \in \mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})} \sum_{\tilde{x}_k^n \in \mathcal{T}_{\tilde{V}^{n-k}}(\tilde{y}_k^n)} \exp\{-(n-l+1)R_x - (n-k+1)R_y\}\Big] p_{xy}(x^n,y^n) \quad (80)$$

In (80) we enumerate all the source sequences in a way that allows us to focus on the types of the important subsequences. We enumerate the possibly misleading candidate sequences in terms of their suffixes types. We restrict the sum to those pairs $(\tilde{x}^n,\tilde{y}^n)$ that could lead to mistaken decoding, defining the compact notation $S(P^{n-k},P^{k-l},V^{n-k},V^{k-l}) \triangleq (k-l)H(V^{k-l}|P^{k-l}) + (n-k+1)H(P^{n-k} \times V^{n-k})$, which is the weighted suffix entropy condition rewritten in terms of types.

Note that the summations within the minimization in (80) do not depend on the arguments within these sums. Thus, we can bound this sum separately to get a bound on the number of possibly misleading source pairs $(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$.

$$\sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k} \text{ s.t.} \\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l}) < \\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} \sum_{\tilde{y}_k^n \in \mathcal{T}_{\tilde{P}^{n-k}}} \sum_{\tilde{x}_l^{k-1} \in \mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})} \sum_{\tilde{x}_k^n \in \mathcal{T}_{\tilde{V}^{n-k}}(\tilde{y}_k^n)}$$

$$\leq \sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k} \text{ s.t.} \\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l}) < \\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} \sum_{\tilde{y}_k^n \in \mathcal{T}_{\tilde{P}^{n-k}}} |\mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})||\mathcal{T}_{\tilde{V}^{n-k}}(\tilde{y}_k^n)| \quad (81)$$

$$\leq \sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k} \text{ s.t.} \\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l}) < \\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} |\mathcal{T}_{\tilde{P}^{n-k}}| \exp\{(k-l)H(\tilde{V}^{k-l}|P^{k-l})\} \exp\{(n-k+1)H(\tilde{V}^{n-k}|\tilde{P}^{n-k})\} \quad (82)$$

$$\leq \sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k} \text{ s.t.} \\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l}) < \\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} \exp\{(k-l)H(\tilde{V}^{k-l}|P^{k-l}) + (n-k+1)H(\tilde{P}^{n-k} \times \tilde{V}^{n-k})\} \quad (83)$$

$$\leq \sum_{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k}} \exp\{(k-l)H(V^{k-l}|P^{k-l}) + (n-k+1)H(P^{n-k} \times V^{n-k})\} \quad (84)$$

$$\leq (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \exp\{(k-l)H(V^{k-l}|P^{k-l}) + (n-k+1)H(P^{n-k} \times V^{n-k})\} \quad (85)$$

In (81) we sum over all $\tilde{x}_l^{k-1} \in \mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})$. In (82) we use standard bounds, e.g., $|\mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})| \le \exp\{(k-l)H(\tilde{V}^{k-l}|P^{k-l})\}$ since $y_l^{k-1} \in \mathcal{T}_{P^{k-l}}$. We also sum over all $\tilde{x}_k^n \in \mathcal{T}_{\tilde{V}^{n-k}}(\tilde{y}_k^n)$ and over all $\tilde{y}_k^n \in \mathcal{T}_{\tilde{P}^{n-k}}$ in (82). By definition of the decoding rule $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ can only lead to a decoding error if $(k-l)H(\tilde{V}^{k-l}|P^{k-l})] + (n-k+1)H(\tilde{P}^{n-k} \times \tilde{V}^{n-k}) < (k-l)H(V^{k-l}|P^{k-l}) + (n-k+1)H(P^{n-k} \times V^{n-k})$. In (85) we apply the polynomial bound on the number of types.

We substitute (85) into (80) and pull out the polynomial term, giving

$$
\begin{aligned}
p_n(l,k) \le\ & (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}} \sum_{\substack{y_l^{k-1} \in \mathcal{T}_{P^{k-l}}, \\ y_k^n \in \mathcal{T}_{P^{n-k}}}} \sum_{\substack{x_l^{k-1} \in \mathcal{T}_{V^{k-l}(y_l^{k-1})}, \\ x_k^n \in \mathcal{T}_{V^{n-k}(y_k^n)}}} \\
& \min\Big[1, \exp\{-(k-l)[R_x - H(V^{k-l}|P^{k-l})] - (n-k+1)[R_x + R_y - H(V^{n-k} \times P^{n-k})]\}\Big]\, p_{x_l^n, y_l^n}(x_l^n, y_l^n) \\
\le & (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}} \\
& \exp\Big\{\max\Big[0, -(k-l)[R_x - H(V^{k-l}|P^{k-l})] - (n-k+1)[R_x + R_y - H(V^{n-k} \times P^{n-k})]\Big]\Big\} \\
& \exp\big\{-(k-l)D(V^{k-l} \times P^{k-l} \| p_{xy}) - (n-k+1)D(V^{n-k} \times P^{n-k} \| p_{xy})\big\} \qquad (86) \\
\le & (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}} \exp\Big\{-(n-l+1)\Big[\lambda D(V^{k-l} \times P^{k-l} \| p_{xy}) + \bar{\lambda} D(V^{n-k} \times P^{n-k} \| p_{xy}) \\
& + \big|\lambda[R_x - H(V^{k-l}|P^{k-l})] + \bar{\lambda}[R_x + R_y - H(V^{n-k} \times P^{n-k})]\big|^+\Big]\Big\} \qquad (87) \\
\le & (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}} \exp\Big\{-(n-l+1) \inf_{\tilde{\mathsf{x}},\tilde{\mathsf{y}},\bar{\mathsf{x}},\bar{\mathsf{y}}} \Big[\lambda D(p_{\tilde{\mathsf{x}},\tilde{\mathsf{y}}} \| p_{xy}) + \bar{\lambda} D(p_{\bar{\mathsf{x}},\bar{\mathsf{y}}} \| p_{xy}) \\
& + \big|\lambda[R_x - H(\tilde{\mathsf{x}}|\tilde{\mathsf{y}})] + \bar{\lambda}[R_x + R_y - H(\bar{\mathsf{x}},\bar{\mathsf{y}})]\big|^+\Big]\Big\} \qquad (88) \\
\le & (n-l+2)^{4|\mathcal{X}||\mathcal{Y}|} \exp\{-(n-l+1)E_x(R_x,R_y,\lambda)\} \le K_1 \exp\{-(n-l+1)[E_x(R_x,R_y,\lambda) - \gamma]\} \qquad (89) \\
& \qquad (90)
\end{aligned}
$$

In (86) we use the memoryless property of the source, and exponential bounds on the probability of observing (x_l^{k-1}, y_l^{k-1}) and (x_k^n, y_k^n). In (87) we pull out $(n-l+1)$ from all terms, noticing that $\lambda = (k-l)/(n-l+1) \in [0,1]$ and $\bar{\lambda} \triangleq 1 - \lambda = (n-k+1)/(n-l+1)$. In (88) we minimize the exponent over all choices of distributions $p_{\tilde{\mathsf{x}},\tilde{\mathsf{y}}}$ and $p_{\bar{\mathsf{x}},\bar{\mathsf{y}}}$. In (89) we define the universal random coding exponent $E_x(R_x,R_y,\lambda) \triangleq \inf_{\tilde{\mathsf{x}},\tilde{\mathsf{y}},\bar{\mathsf{x}},\bar{\mathsf{y}}}\{\lambda D(p_{\tilde{\mathsf{x}},\tilde{\mathsf{y}}} \| p_{xy}) + \bar{\lambda} D(p_{\bar{\mathsf{x}},\bar{\mathsf{y}}} \| p_{xy}) + \big|\lambda[R_x - H(\tilde{\mathsf{x}}|\tilde{\mathsf{y}})] + \bar{\lambda}[R_x + R_y - H(\bar{\mathsf{x}},\bar{\mathsf{y}})]\big|^+\}$ where $0 \le \lambda \le 1$ and $\bar{\lambda} = 1 - \lambda$. We also incorporate the number of conditional and marginal types into the polynomial bound, as well as the sum over k, and then push the polynomial into the exponent since for any polynomial F, $\forall E, \epsilon > 0$, there exists $C > 0$, s.t. $F(\Delta)e^{-\Delta E} \le Ce^{-\Delta(E-\epsilon)}$. ■

A similar derivation yields a bound on $p_n(l,k)$ for $l \ge k$.

Combining Lemmas 6 and 5, and then following the same derivation for ML decoding yields Theorem 7.

VIII. Future Directions

A. Stationary-ergodic sources and universality

[12] extends the block-coding proofs to the Slepian-Wolf problem for stationary-ergodic sources using AEP arguments. To have a similar extension to the streaming context, possibly additional regularity conditions will be required so that error exponents can be achieved. To achieve universality over sources, it is possible that further technical restrictions will be required. For the case of distributed Markov sources however, it seems quite clear that all the arguments in this paper will easily generalize. In that case, following the approach we take in [13], the source can be "segmented" into small blocks and the endpoints[15] of the blocks can be encoded perfectly at essentially zero rate. Conditioned on these endpoints, the blocks are then iid, with the endpoints representing a third stream of perfectly known side-information.

[15]For a Markov source of known order k, the endpoint is just k successive symbols at the end of the block.

B. Upper bounds and demonstrating optimal delays

This paper dealt entirely with achievability of certain error exponents. Ideally, we would have corresponding upper bounds demonstrating that no higher exponents are possible. In the block-coding case, problem 3.7.1 in [8] provides a simple upper-bound. However, the nature of the error exponents in the streaming case might be more complicated. [6] provides an upper bound and matching achievable scheme for point-to-point source-coding with delay and this bound extends naturally to the case where side-information is known at both the encoder and the decoder. [14] provides an upper bound for the case of side-information known only at the decoder, and this bound is tight for certain symmetric cases. However, both of these extended single encoder arguments from [15] that do not immediately generalize to the case of multiple encoders.

C. Trading off error exponents for the different source terminals

For multiple terminal systems, different error exponents can be achieved for different users or sources. For channel coding, the encoders can choose different distributions while generating the randomized code book to achieve an error exponent trade-off among different users. In [16], the error exponent region is studied for the Gaussian multiple access channel and the broadcast channel within the block-coding paradigm. It is unclear whether similar tradeoffs are possible within the streaming Slepian Wolf problems considered here since there is nothing immediately comparable to the flexibility we have in choosing the "input distribution" for channel coding problems.

D. Adaptation and limited feedback

An interesting extension is to adaptive universal streaming Slepian Wolf encoders. The decoders we use in this paper are based on empirical statistics. Therefore they can be used even if source statistics are unknown. The current proposal will work regardless of source and side information statistics as long as the conditional entropy $H(x|y)$ is less than the encoding rate. Even if there is uncertainty in statistics, the anytime nature of the coding system should enable the system to adapt on-line to the unknown entropy rate if some feedback channel is available. The feedback channel would be used to order increases (or decreases) in the binning rate. An increase (or decrease) could be triggered by examining the difference between two quantities: the minimal empirical joint entropy between the decoded sequence and observation, and the empirical joint entropy between the particular sequence and observation yielding the second-lowest joint entropy. If there is a large difference between these two entropies, we are using rate excessively, and the rate of communication can be reduced. If the difference is negligible, then it's likely we are not decoding correctly. Our target should be to keep this difference at roughly ϵ. In the current context, this is analogous to the rate margin by which we choose to exceed the known conditional entropy.

ACKNOWLEDGMENTS

The authors wish to acknowledge a desire expressed by Zixiang Xiong and subsequent hallway discussions during ITW 2004 that helped precipitate the current line of research. This work was supported in part by NSF ITR Grant No. CNS-0326503.

APPENDIX

In this section we show that the maximum likelihood (ML) error exponent equals the universal error exponent. We show that for all γ,

$$E_x^{ML}(R_x, R_y, \gamma) = E_x^{UN}(R_x, R_y, \gamma)$$

Where the ML error exponent:

$$\begin{aligned} E_x^{ML}(R_x, R_y, \gamma) &= \sup_{\rho \in [0,1]} \{\gamma E_{x|y}(R_x, \rho) + (1-\gamma) E_{xy}(R_x, R_y, \rho)\} \\ &= \sup_{\rho \in [0,1]} \{\rho R^{(\gamma)} - \gamma \log(\sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho}) - (1-\gamma)(1+\rho)\log(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})\} \\ &= \sup_{\rho \in [0,1]} \{E_x^{ML}(R_x, R_y, \gamma, \rho)\} \end{aligned}$$

Write the function inside the sup argument as $E_x^{ML}(R_x, R_y, \gamma, \rho)$. The universal error exponent:

$$\begin{aligned} E_x^{UN}(R_x, R_y, \gamma) &= \inf_{q_{xy}, o_{xy}} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) \\ &\quad + \max\{0, \gamma(R_x - H(q_{x|y})) + (1-\gamma)(R_x + R_y - H(o_{xy}))\}\} \\ &= \inf_{q_{xy}, o_{xy}} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) + \max\{0, R^{(\gamma)} - \gamma H(q_{x|y}) - (1-\gamma) H(o_{xy})\}\} \end{aligned}$$

Here we define $R^{(\gamma)} = \gamma R_x + (1-\gamma)(R_x + R_y) > \gamma H(p_{x|y}) + (1-\gamma) H(p_{xy})$. For notational simplicity, we write q_{xy} and o_{xy} as two arbitrary joint distributions on $\mathcal{X} \times \mathcal{Y}$ instead of $p_{\bar{x}\bar{y}}$ and $p_{\bar{\bar{x}}\bar{\bar{y}}}$. We still write p_{xy} as the distribution of the source.

Before the proof, we define a pair of distributions that we will need.

Definition 4: Tilted distribution of p_{xy}: p_{xy}^{ρ}, for all $\rho \in [-1, \infty)$

$$p_{xy}^{\rho}(x,y) = \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}}$$

The entropy of the tilted distribution is written as $H(p_{xy}^{\rho})$. Obviously $p_{xy}^{0} = p_{xy}$.

Definition 5: $x-y$ tilted distribution of p_{xy}: $\bar{p}_{xy}^{\rho}$, for all $\rho \in [-1, +\infty)$

$$\begin{aligned} \bar{p}_{xy}^{\rho}(x,y) &= \frac{[\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}]^{1+\rho}}{\sum_t [\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}]^{1+\rho}} \times \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}} \\ &= \frac{A(y,\rho)}{B(\rho)} \times \frac{C(x,y,\rho)}{D(y,\rho)} \end{aligned}$$

Where

$$\begin{aligned} A(y,\rho) &= [\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}]^{1+\rho} = D(y,\rho)^{1+\rho} \\ B(\rho) &= \sum_s [\sum_t p_{xy}(s,t)^{\frac{1}{1+\rho}}]^{1+\rho} = \sum_y A(y,\rho) \\ C(x,y,\rho) &= p_{xy}(x,y)^{\frac{1}{1+\rho}} \\ D(y,\rho) &= \sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}} = \sum_x C(x,y,\rho) \end{aligned}$$

The marginal distribution for y is $\frac{A(y,\rho)}{B(\rho)}$. Obviously $\bar{p}_{xy}^{0} = p_{xy}$. Write the conditional distribution of x given y under distribution $\bar{p}_{xy}^{\rho}$ as $\bar{p}_{x|y}^{\rho}$, where $\bar{p}_{x|y}^{\rho}(x,y) = \frac{C(x,y,\rho)}{D(y,\rho)}$, and the conditional entropy of x given y under distribution $\bar{p}_{xy}^{\rho}$ as $H(\bar{p}_{x|y}^{\rho})$. Obviously $H(\bar{p}_{x|y}^{0}) = H(p_{x|y})$.
The conditional entropy of x given y for the $x-y$ tilted distribution is

$$H(\bar{p}_{x|y=y}^{\rho}) = -\sum_x \frac{C(x,y,\rho)}{D(y,\rho)} \log(\frac{C(x,y,\rho)}{D(y,\rho)})$$

We introduce $A(y,\rho)$, $B(\rho)$, $C(x,y,\rho)$, $D(y,\rho)$ to simplify the notations. Some of their properties are shown in Lemma 10.

While tilted distributions are common optimal distributions in large deviation theory, it is useful to contemplate why we need to introduce these *two* tilted distributions. In the proof of Theorem 8, through a Lagrange multiplier argument, we will show that $\{p_{xy}^{\rho} : \rho \in [-1, +\infty)\}$ is the family of distributions that minimize the Kullback−Leibler distance to p_{xy} with fixed *entropy* and $\{\bar{p}_{xy}^{\rho} : \rho \in [-1, +\infty)\}$ is the family of distributions that minimize the Kullback−Leibler distance to p_{xy} with fixed *conditional entropy*. Using a Lagrange multiplier argument, we parametrize the universal error exponent $E_x^{UN}(R_x, R_y, \gamma)$ in terms of ρ and show the equivalence of the universal and maximum likelihood error exponents.

Now we are ready to prove Theorem 8: $E_x^{ML}(R_x, R_y, \gamma) = E_x^{UN}(R_x, R_y, \gamma)$.
Proof:

A. case 1: $\gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}) < R^{(\gamma)} < \gamma H(\bar{p}^1_{x|y}) + (1-\gamma)H(p^1_{xy})$.

First, from Lemma 16 and Lemma 17:

$$\frac{\partial E_x^{ML}(R_x, R_y, \gamma, \rho)}{\partial \rho} = R^{(\gamma)} - \gamma H(\bar{p}^\rho_{x|y}) - (1-\gamma)H(p^\rho_{xy})$$

Then, using Lemma 7 and Lemma 11, we have:

$$\frac{\partial^2 E_x^{ML}(R_x, R_y, \gamma, \rho)}{\partial \rho} \leq 0$$

.

So ρ maximize $E_x^{ML}(R_x, R_y, \gamma, \rho)$, if and only if:

$$0 = \frac{\partial E_x^{ML}(R_x, R_y, \gamma, \rho)}{\partial \rho} = R^{(\gamma)} - \gamma H(\bar{p}^\rho_{x|y}) - (1-\gamma)H(p^\rho_{xy}) \tag{91}$$

Because $R^{(\gamma)}$ is in the interval $[\gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}), \gamma H(\bar{p}^1_{x|y}) + (1-\gamma)H(p^1_{xy})]$ and the entropy functions monotonically-increase over ρ, we can find $\rho^* \in (0,1)$, s.t.

$$\gamma H(\bar{p}^{\rho^*}_{x|y}) + (1-\gamma)H(p^{\rho^*}_{xy}) = R^{(\gamma)}$$

Using Lemma 14 and Lemma 15 we get:

$$E_x^{ML}(R_x, R_y, \gamma) \quad = \quad \gamma D(\bar{p}^{\rho^*}_{xy} \| p_{xy}) + (1-\gamma) D(p^{\rho^*}_{xy} \| p_{xy}) \tag{92}$$

Where $\gamma H(\bar{p}^{\rho^*}_{x|y}) + (1-\gamma)H(p^{\rho^*}_{xy}) = R^{(\gamma)}$, ρ^* is generally unique because both $H(\bar{p}^\rho_{x|y})$ and $H(p^\rho_{xy})$ are strictly increasing with ρ.

Secondly

$$\begin{aligned}
& E_x^{UN}(R_x, R_y, \gamma) \\
= \quad & \inf_{q_{xy}, o_{xy}} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) + \max\{0, R^{(\gamma)} - \gamma H(q_{x|y}) - (1-\gamma)H(o_{xy})\}\} \\
= \quad & \inf_b \{ \inf_{q_{xy}, o_{xy}: \gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) = b} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) + \max(0, R^{(\gamma)} - b)\}\} \\
= \quad & \inf_{b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})} \{ \inf_{q_{xy}, o_{xy}: \gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) = b} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) \\
& + \max(0, R^{(\gamma)} - b)\}\}
\end{aligned} \tag{93}$$

The last equality is true because, for $b < \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}) < R^{(\gamma)}$,

$$\begin{aligned}
& \inf_{q_{xy}, o_{xy}: \gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) = b} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) + \max(0, R^{(\gamma)} - b)\}\} \\
\geq \quad & 0 + R^{(\gamma)} - b \\
= \quad & \inf_{q_{xy}, o_{xy}: H(q_{x|y}) = H(p_{x|y}), H(o_{xy}) = H(p_{xy})} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) + \max(0, R^{(\gamma)} - b)\}\} \\
\geq \quad & \inf_{q_{xy}, o_{xy}: H(q_{x|y}) = H(p_{x|y}), H(o_{xy}) = H(p_{xy})} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) \\
& + \max(0, R^{(\gamma)} - \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}))\}\} \\
\geq \quad & \inf_{q_{xy}, o_{xy}: \gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) = \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) \\
& + \max(0, R^{(\gamma)} - \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}))\}\}
\end{aligned}$$

Fixing $b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})$, the inner infimum in (93) is an optimization problem on q_{xy}, o_{xy} with equality constraints $\sum_x \sum_y q_{xy}(x,y) = 1$, $\sum_x \sum_y o_{xy}(x,y) = 1$ and $\gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) = b$ and the obvious inequality constraints $0 \leq q_{xy}(x,y) \leq 1, 0 \leq o_{xy}(x,y) \leq 1, \forall x, y$. In the following formulation of the optimization problem, we relax one equality constraint to an inequality constraint $\gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) \geq b$ to make the optimization problem $convex$. It turns out later that the optimal solution to the relaxed problem is also the optimal solution to the original problem because $b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})$. The resulting optimization problem is:

$$\begin{aligned}
&\inf_{q_{xy}, o_{xy}} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy})\} \\
&\text{s.t.} \sum_x \sum_y q_{xy}(x,y) = 1 \\
&\sum_x \sum_y o_{xy}(x,y) = 1 \\
&b - \gamma H(q_{x|y}) - (1-\gamma)H(o_{xy}) \leq 0 \\
&0 \leq q_{xy}(x,y) \leq 1, \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y} \\
&0 \leq o_{xy}(x,y) \leq 1, \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y}
\end{aligned} \tag{94}$$

The above optimization problem is *convex* because the objective function and the inequality constraint functions are convex and the equality constraint functions are affine[17]. The Lagrange multiplier function for this convex optimization problem is:

$$\begin{aligned}
& L(q_{xy}, o_{xy}, \rho, \mu_1, \mu_2, \nu_1, \nu_2, \nu_3, \nu_4) \\
= \; & \gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy}) \\
& + \mu_1 (\sum_x \sum_y q_{xy}(x,y) - 1) + \mu_2 (\sum_x \sum_y o_{xy}(x,y) - 1) \\
& + \rho(b - \gamma H(q_{x|y}) - (1-\gamma)H(o_{xy})) \\
& + \sum_x \sum_y \left\{ \nu_1(x,y)(-q_{xy}(x,y)) + \nu_2(x,y)(1 - q_{xy}(x,y)) + \nu_3(x,y)(-o_{xy}(x,y)) + \nu_4(x,y)(1 - o_{xy}(x,y)) \right\}
\end{aligned} \tag{95}$$

Where ρ, μ_1, μ_2 are real numbers and $\nu_i \in R^{|\mathcal{X}||\mathcal{Y}|}$, $i = 1, 2, 3, 4$.

According to the KKT conditions for convex optimization[17], q_{xy}, o_{xy} minimize the convex optimization problem in (94) if and only if the following conditions are simultaneously satisfied for some q_{xy}, o_{xy}, μ_1, μ_2, ν_1, ν_2, ν_3, ν_4 and ρ:

$$\begin{aligned}
0 \; &= \; \frac{\partial L(q_{xy}, o_{xy}, \rho, \mu_1, \mu_2, \nu_1, \nu_2, \nu_3, \nu_4)}{\partial q_{xy}(x,y)} \\
&= \; \gamma[-\log(p_{xy}(x,y)) + (1+\rho)(1 + \log(q_{xy}(x,y))) + \rho \log(\sum_s q_{xy}(s,y))] + \mu_1 - \nu_1(x,y) - \nu_2(x,y) \\
0 \; &= \; \frac{\partial L(q_{xy}, o_{xy}, \rho, \mu_1, \mu_2, \nu_1, \nu_2, \nu_3, \nu_4)}{\partial o_{xy}(x,y)} \\
&= \; (1-\gamma)[-\log(p_{xy}(x,y)) + (1+\rho)(1 + \log(o_{xy}(x,y)))] + \mu_2 - \nu_3(x,y) - \nu_4(x,y)
\end{aligned} \tag{96}$$

For all x, y and

$$\begin{aligned}
&\sum_x \sum_y q_{xy}(x,y) = 1 \\
&\sum_x \sum_y o_{xy}(x,y) = 1 \\
&\rho(\gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) - b) = 0 \\
&\rho \geq 0 \\
&\nu_1(x,y)(-q_{xy}(x,y)) = 0, \quad \nu_2(x,y)(1-q_{xy}(x,y)) = 0 \quad \forall x,y \\
&\nu_3(x,y)(-o_{xy}(x,y)) = 0, \quad \nu_4(x,y)(1-o_{xy}(x,y)) = 0 \quad \forall x,y \\
&\nu_i(x,y) \geq 0, \quad \forall x,y,i = 1,2,3,4
\end{aligned} \tag{97}$$

Solving the above standard Lagrange multiplier equations (96) and (97), we have:

$$\begin{aligned}
q_{xy}(x,y) &= \frac{[\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho_b}}]^{1+\rho_b}}{\sum_t [\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho_b}}]^{1+\rho_b}} \frac{p_{xy}(x,y)^{\frac{1}{1+\rho_b}}}{\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho_b}}} \\
&= \bar{p}_{xy}^{\rho_b}(x,y) \\
o_{xy}(x,y) &= \frac{p_{xy}(x,y)^{\frac{1}{1+\rho_b}}}{\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho_b}}} \\
&= p_{xy}^{\rho_b}(x,y) \\
\nu_i(x,y) &= 0 \quad \forall x,y,i = 1,2,3,4 \\
\rho &= \rho_b
\end{aligned} \tag{98}$$

Where ρ_b satisfies the following condition

$$\gamma H(\bar{p}_{x|y}^{\rho_b}) + (1-\gamma)H(p_{xy}^{\rho_b}) = b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})$$

and thus $\rho_b \geq 0$ because both $H(\bar{p}_{x|y}^{\rho})$ and $H(p_{xy}^{\rho})$ are monotonically increasing with ρ as shown in Lemma 7 and Lemma 11.

Notice that all the KKT conditions are simultaneously satisfied with the inequality constraint $\gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) \geq b$ being met with equality. Thus, the relaxed optimization problem has the same optimal solution as the original problem as promised. The optimal q_{xy} and o_{xy} are the $x-y$ tilted distribution $\bar{p}_{xy}^{\rho_b}$ and standard tilted distribution $p_{xy}^{\rho_b}$ of p_{xy} with the same parameter $\rho_b \geq 0$. chosen s.t.

$$\gamma H(\bar{p}_{x|y}^{\rho_b}) + (1-\gamma)H(p_{xy}^{\rho_b}) = b$$

Now we have :

$$\begin{aligned}
& E_x^{UN}(R_x, R_y, \gamma) \\
= & \inf_{b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})} \{ \inf_{q_{xy}, o_{xy}: \gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) = b} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy}) + \max(0, R^{(\gamma)} - b)\}\} \\
= & \inf_{b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})} \{\gamma D(\bar{p}_{xy}^{\rho_b}||p_{xy}) + (1-\gamma)D(p_{xy}^{\rho_b}||p_{xy}) + \max(0, R^{(\gamma)} - b)\} \\
= & \min[\inf_{\rho \geq 0: R^{(\gamma)} \geq \gamma H(\bar{p}_{x|y}^{\rho}) + (1-\gamma)H(p_{xy}^{\rho})} \{\gamma D(\bar{p}_{xy}^{\rho}||p_{xy}) + (1-\gamma)D(p_{xy_\rho}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{\rho}) - (1-\gamma)H(p_{xy}^{\rho})\}, \\
& \inf_{\rho \geq 0: R^{(\gamma)} \leq \gamma H(\bar{p}_{x|y}^{\rho}) + (1-\gamma)H(p_{xy}^{\rho})} \{\gamma D(\bar{p}_{xy}^{\rho}||p_{xy}) + (1-\gamma)D(p_{xy_\rho}||p_{xy})\}]
\end{aligned} \tag{99}$$

Notice that $H(p_{xy}^{\rho})$, $H(\bar{p}_{x|y}^{\rho})$, $D(\bar{p}_{xy}^{\rho}||p_{xy})$ and $D(p_{xy}^{\rho}||p_{xy})$ are all strictly increasing with $\rho > 0$ as shown in Lemma 11, Lemma 12, Lemma 7 and Lemma 8 later in this appendix. We have:

$$\begin{aligned}
& \inf_{\rho \geq 0: R^{(\gamma)} \leq \gamma H(\bar{p}_{x|y}^{\rho}) + (1-\gamma)H(p_{xy}^{\rho})} \{\gamma D(\bar{p}_{xy}^{\rho}||p_{xy}) + (1-\gamma)D(p_{xy}^{\rho}||p_{xy})\} \\
= & \gamma D(\bar{p}_{xy}^{\rho^*}||p_{xy}) + (1-\gamma)D(p_{xy}^{\rho^*}||p_{xy})
\end{aligned} \tag{100}$$

where $R^{(\gamma)} = \gamma H(\bar{p}^{\rho^*}_{x|y}) + (1-\gamma)H(p^{\rho^*}_{xy})$. Applying the results in Lemma 13 and Lemma 9, we get:

$$\begin{aligned}
&\inf_{\rho\geq 0: R^{(\gamma)} \geq \gamma H(\bar{p}^{\rho}_{x|y}) + (1-\gamma)H(p^{\rho}_{xy})} \{\gamma D(\bar{p}^{\rho}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho}_{xy}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}^{\rho}_{x|y}) - (1-\gamma)H(p^{\rho}_{xy})\} \\
&= \gamma D(\bar{p}^{\rho}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho}_{xy}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}^{\rho}_{x|y}) - (1-\gamma)H(p^{\rho}_{xy})|_{\rho=\rho^*} \\
&= \gamma D(\bar{p}^{\rho^*}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho^*}_{xy}||p_{xy}) \qquad (101)
\end{aligned}$$

This is true because for $\rho : R^{(\gamma)} \geq \gamma H(\bar{p}^{\rho}_{x|y}) + (1-\gamma)H(p^{\rho}_{xy})$, we know $\rho \leq 1$ because of the range of $R^{(\gamma)}$: $R^{(\gamma)} < \gamma H(\bar{p}^{1}_{x|y}) + (1-\gamma)H(p^{1}_{xy})$. Substituting (100) and (101) into (99), we get

$$\begin{aligned}
E_x^{UN}(R_x, R_y, \gamma) \quad &= \quad \gamma D(\bar{p}^{\rho^*}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho^*}_{xy}||p_{xy}) \\
&\quad \text{where} \quad R^{(\gamma)} = \gamma H(\bar{p}^{\rho^*}_{x|y}) + (1-\gamma)H(p^{\rho^*}_{xy}) \qquad (102)
\end{aligned}$$

So for $\gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}) \leq R^{(\gamma)} \leq \gamma H(\bar{p}^{1}_{x|y}) + (1-\gamma)H(p^{1}_{xy})$, from (92) we have the desired property:

$$E_x^{ML}(R_x, R_y, \gamma) = E_x^{UN}(R_x, R_y, \gamma)$$

B. case 2: $R^{(\gamma)} \geq \gamma H(\bar{p}^{1}_{x|y}) + (1-\gamma)H(p^{1}_{xy})$.

In this case, for all $0 \leq \rho \leq 1$

$$\frac{\partial E_x^{ML}(R_x, R_y, \gamma, \rho)}{\partial \rho} = R^{(\gamma)} - \gamma H(\bar{p}^{\rho}_{x|y}) - (1-\gamma)H(p^{\rho}_{xy}) \geq R^{(\gamma)} - \gamma H(\bar{p}^{1}_{x|y}) - (1-\gamma)H(p^{1}_{xy}) \geq 0$$

So ρ takes value 1 to maximize the error exponent $E_x^{ML}(R_x, R_y, \gamma, \rho)$, thus

$$E_x^{ML}(R_x, R_y, \gamma) = R^{(\gamma)} - \gamma \log(\sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{2}})^2) - 2(1-\gamma)\log(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{2}}) \qquad (103)$$

Using the same convex optimization techniques as case A, we notice the fact that $\rho^* \geq 1$ for $R^{(\gamma)} = \gamma H(\bar{p}^{\rho^*}_{x|y}) + (1-\gamma)H(p^{\rho^*}_{xy})$. Then applying Lemma 13 and Lemma 9, we have:

$$\begin{aligned}
&\inf_{\rho\geq 0: R^{(\gamma)} \geq \gamma H(\bar{p}^{\rho}_{x|y}) + (1-\gamma)H(p^{\rho}_{xy})} \{\gamma D(\bar{p}^{\rho}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho}_{xy}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}^{\rho}_{x|y}) - (1-\gamma)H(p_{xy_\rho})\}, \\
&= \gamma D(\bar{p}^{1}_{xy}||p_{xy}) + (1-\gamma)D(p^{1}_{xy}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}^{1}_{x|y}) - (1-\gamma)H(p^{1}_{xy})
\end{aligned}$$

And

$$\begin{aligned}
&\inf_{\rho\geq 0: R^{(\gamma)} \leq \gamma H(\bar{p}^{\rho}_{x|y}) + (1-\gamma)H(p^{\rho}_{xy})} \{\gamma D(\bar{p}^{\rho}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho}_{xy}||p_{xy})\}] \\
&= \gamma D(\bar{p}^{\rho^*}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho^*}_{xy}||p_{xy}) \\
&= \gamma D(\bar{p}^{\rho^*}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho^*}_{xy}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}^{\rho^*}_{x|y}) - (1-\gamma)H(p^{\rho^*}_{xy}) \\
&\leq \gamma D(\bar{p}^{1}_{xy}||p_{xy}) + (1-\gamma)D(p^{1}_{xy}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}^{1}_{x|y}) - (1-\gamma)H(p^{1}_{xy})
\end{aligned}$$

Finally:

$$
\begin{aligned}
& E_x^{UN}(R_x, R_y, \gamma) \\
= & \inf_{b \geq \gamma H(p_{x|y}) + (1-\gamma) H(p_{xy})} \{ \inf_{q_{xy}, o_{xy} : \gamma H(q_{x|y}) + (1-\gamma) H(o_{xy}) = b} \{ \gamma D(q_{xy} || p_{xy}) + (1-\gamma) D(o_{xy} || p_{xy}) + \max(0, R^{(\gamma)} - b) \} \} \\
= & \inf_{b \geq \gamma H(p_{x|y}) + (1-\gamma) H(p_{xy})} \{ \gamma D(\bar{p}_{xy}^{\rho_b} || p_{xy}) + (1-\gamma) D(p_{xy}^{\rho_b} || p_{xy}) + \max(0, R^{(\gamma)} - b) \} \\
= & \min[\inf_{\rho \geq 0 : R^{(\gamma)} \geq \gamma H(\bar{p}_{x|y}^{\rho}) + (1-\gamma) H(p_{xy}^{\rho})} \{ \gamma D(\bar{p}_{xy}^{\rho} || p_{xy}) + (1-\gamma) D(p_{xy}^{\rho} || p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{\rho}) - (1-\gamma) H(p_{xy}^{\rho}) \}, \\
& \inf_{\rho \geq 0 : R^{(\gamma)} \leq \gamma H(\bar{p}_{x|y}^{\rho}) + (1-\gamma) H(p_{xy}^{\rho})} \{ \gamma D(\bar{p}_{xy}^{\rho} || p_{xy}) + (1-\gamma) D(p_{xy}^{\rho} || p_{xy}) \}] \\
= & \gamma D(\bar{p}_{xy}^{1} || p_{xy}) + (1-\gamma) D(p_{xy}^{1} || p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{1}) - (1-\gamma) H(p_{xy}^{1}) \\
= & R^{(\gamma)} - \gamma \log(\sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{2}})^2) - 2(1-\gamma) \log(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{2}}) \qquad (104)
\end{aligned}
$$

The last equality is true by setting $\rho = 1$ in Lemma 14 and Lemma 15.

Again, $E_x^{ML}(R_x, R_y, \gamma) = E_x^{UN}(R_x, R_y, \gamma)$, thus we finish the proof. ■

C. Technical Lemmas

Some technical lemmas we used in the above proof of Theorem 8 are now discussed:

Lemma 7: $\frac{\partial H(p_{xy}^{\rho})}{\partial \rho} \geq 0$

Proof: From the definition of the tilted distribution we have the following observation:

$\log(p_{xy}^{\rho}(x_1, y_1)) - \log(p_{xy}^{\rho}(x_2, y_2)) = \log(p_{xy}(x_1, y_1)^{\frac{1}{1+\rho}}) - \log(p_{xy}(x_2, y_2)^{\frac{1}{1+\rho}})$

Using the above equality, we first derive the derivative of the tilted distribution, for all x, y

$$
\begin{aligned}
\frac{\partial p_{xy}^{\rho}(x,y)}{\partial \rho} &= \frac{-1}{(1+\rho)^2} \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}} \log(p_{xy}(x,y)) (\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}})}{(\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}})^2} \\
&\quad - \frac{-1}{(1+\rho)^2} \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}} (\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}} \log(p_{xy}(s,t)))}{(\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}})^2} \\
&= \frac{-1}{1+\rho} p_{xy}^{\rho}(x,y) [\log(p_{xy}(x,y)^{\frac{1}{1+\rho}}) - \sum_t \sum_s p_{xy}^{\rho}(s,t) \log(p_{xy}(s,t)^{\frac{1}{1+\rho}})] \\
&= \frac{-1}{1+\rho} p_{xy}^{\rho}(x,y) [\log(p_{xy}^{\rho}(x,y)) - \sum_t \sum_s p_{xy}^{\rho}(s,t) \log(p_{xy}^{\rho}(s,t))] \\
&= - \frac{p_{xy}^{\rho}(x,y)}{1+\rho} [\log(p_{xy}^{\rho}(x,y)) + H(p_{xy}^{\rho})] \qquad (105)
\end{aligned}
$$

Then:

$$\begin{aligned}
\frac{\partial H(p^{\rho}_{xy})}{\partial \rho} &= -\frac{\partial \sum_{x,y} p^{\rho}_{xy}(x,y)\log(p^{\rho}_{xy}(x,y))}{\partial \rho} \\
&= -\sum_{x,y}(1+\log(p^{\rho}_{xy}(x,y)))\frac{\partial p^{\rho}_{xy}(x,y)}{\partial \rho} \\
&= \sum_{x,y}(1+\log(p^{\rho}_{xy}(x,y)))\frac{p^{\rho}_{xy}(x,y)}{1+\rho}(\log(p^{\rho}_{xy}(x,y))+H(p^{\rho}_{xy})) \\
&= \frac{1}{1+\rho}\sum_{x,y} p^{\rho}_{xy}(x,y)\log(p^{\rho}_{xy}(x,y))(\log(p^{\rho}_{xy}(x,y))+H(p^{\rho}_{xy})) \\
&= \frac{1}{1+\rho}[\sum_{x,y} p^{\rho}_{xy}(x,y)(\log(p^{\rho}_{xy}(x,y)))^2 - H(p^{\rho}_{xy})^2] \\
&= \frac{1}{1+\rho}[\sum_{x,y} p^{\rho}_{xy}(x,y)(\log(p^{\rho}_{xy}(x,y)))^2 \sum_{x,y} p^{\rho}_{xy}(x,y) - H(p^{\rho}_{xy})^2] \\
&\geq_{(a)} \frac{1}{1+\rho}[(\sum_{x,y} p^{\rho}_{xy}(x,y)\log(p^{\rho}_{xy}(x,y)))^2 - H(p^{\rho}_{xy})^2] \\
&= 0
\end{aligned} \tag{106}$$

where (a) is true by the Cauchy-Schwartz inequality. ■

Lemma 8: $\frac{\partial D(p^{\rho}_{xy}\|P)}{\partial \rho} = \rho\frac{\partial H(p^{\rho}_{xy})}{\partial \rho}$

Proof: As shown in Lemma 14 and Lemma 16 respectively:

$$D(p^{\rho}_{xy}\|p_{xy}) = \rho H(p^{\rho}_{xy}) - (1+\rho)\log(\sum_{x,y} p_{xy}(x,y)^{\frac{1}{1+\rho}})$$

$$H(p^{\rho}_{xy}) = \frac{\partial (1+\rho)\log(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})}{\partial \rho}$$

We have:

$$\begin{aligned}
\frac{\partial D(p^{\rho}_{xy}\|p_{xy})}{\partial \rho} &= H(p^{\rho}_{xy}) + \rho\frac{\partial H(p^{\rho}_{xy})}{\partial \rho} - \frac{\partial (1+\rho)\log(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})}{\partial \rho} \\
&= H(p^{\rho}_{xy}) + \rho\frac{\partial H(p^{\rho}_{xy})}{\partial \rho} - H(p^{\rho}_{xy}) \\
&= \rho\frac{\partial H(p^{\rho}_{xy})}{\partial \rho}
\end{aligned} \tag{107}$$

■

Lemma 9: $sign\frac{\partial [D(p^{\rho}_{xy}\|p_{xy}) - H(p^{\rho}_{xy})]}{\partial \rho} = sign(\rho - 1)$.

Proof: Combining the results of the previous two lemmas, we have:

$$\frac{\partial D(p^{\rho}_{xy}\|p_{xy}) - H(p^{\rho}_{xy})}{\partial \rho} = (\rho-1)\frac{\partial H(p^{\rho}_{xy})}{\partial \rho} = sign(\rho-1)$$

■

Lemma 10: Properties of $\frac{\partial A(y,\rho)}{\partial \rho}$, $\frac{\partial B(\rho)}{\partial \rho}$, $\frac{\partial C(x,y,\rho)}{\partial \rho}$, $\frac{\partial D(y,\rho)}{\partial \rho}$ and $\frac{\partial H(\bar{p}^{\rho}_{x|y=y})}{\partial \rho}$

First,

$$\begin{aligned}
\frac{\partial C(x,y,\rho)}{\partial \rho} &= \frac{\partial p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\partial \rho} = -\frac{1}{1+\rho} p_{xy}(x,y)^{\frac{1}{1+\rho}} \log(p_{xy}(x,y)^{\frac{1}{1+\rho}}) \\
&= -\frac{C(x,y,\rho)}{1+\rho}\log(C(x,y,\rho)) \\
\frac{\partial D(y,\rho)}{\partial \rho} &= \frac{\partial \sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}}{\partial \rho} = -\frac{1}{1+\rho}\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}} \log(p_{xy}(s,y)^{\frac{1}{1+\rho}}) \\
&= -\frac{\sum_x C(x,y,\rho)\log(C(x,y,\rho))}{1+\rho}
\end{aligned} \tag{108}$$

For a differentiable function $f(\rho)$,

$$\frac{\partial f(\rho)^{1+\rho}}{\partial \rho} = f(\rho)^{1+\rho}\log(f(\rho)) + (1+\rho)f(\rho)^{\rho}\frac{\partial f(\rho)}{\partial \rho}$$

So

$$\begin{aligned}
\frac{\partial A(y,\rho)}{\partial \rho} &= \frac{\partial D(y,\rho)^{1+\rho}}{\partial \rho} = D(y,\rho)^{1+\rho}\log(D(y,\rho)) + (1+\rho)D(y,\rho)^{\rho}\frac{\partial D(y,\rho)}{\partial \rho} \\
&= D(y,\rho)^{1+\rho}(\log(D(y,\rho)) - \sum_x \frac{C(x,y,\rho)}{D(y,\rho)}\log(C(x,y,\rho))) \\
&= D(y,\rho)^{1+\rho}(-\sum_x \frac{C(x,y,\rho)}{D(y,\rho)}\log(\frac{C(x,y,\rho)}{D(y,\rho)})) \\
&= A(y,\rho)H(\bar{p}^{\rho}_{x|y=y}) \\
\frac{\partial B(\rho)}{\partial \rho} &= \sum_y \frac{\partial A(y,\rho)}{\partial \rho} = \sum_y A(y,\rho)H(\bar{p}^{\rho}_{x|y=y}) = B(\rho)\sum_y \frac{A(y,\rho)}{B(\rho)}H(\bar{p}^{\rho}_{x|y=y}) = B(\rho)H(\bar{p}^{\rho}_{x|y})
\end{aligned}$$

And last:

$$\begin{aligned}
&\frac{\partial H(\bar{p}^{\rho}_{x|y=y})}{\partial \rho} \\
&= -\sum_x [\frac{\frac{\partial C(x,y,\rho)}{\partial \rho}}{D(y,\rho)} - \frac{C(x,y,\rho)\frac{\partial D(y,\rho)}{\partial \rho}}{D(y,\rho)^2}][1+\log(\frac{C(x,y,\rho)}{D(y,\rho)})] \\
&= -\sum_x [\frac{-\frac{C(x,y,\rho)}{1+\rho}\log(C(x,y,\rho))}{D(y,\rho)} + \frac{C(x,y,\rho)\frac{\sum_s C(s,y,\rho)\log(C(s,y,\rho))}{1+\rho}}{D(y,\rho)^2}][1+\log(\frac{C(x,y,\rho)}{D(y,\rho)})] \\
&= \frac{1}{1+\rho}\sum_x [\bar{p}^{\rho}_{x|y}(x,y)\log(C(x,y,\rho)) - \bar{p}^{\rho}_{x|y}(x,y)\sum_s \bar{p}^{\rho}_{x|y}(s,y)\log(C(s,y,\rho))][1+\log(\bar{p}^{\rho}_{x|y}(x,y))] \\
&= \frac{1}{1+\rho}\sum_x \bar{p}^{\rho}_{x|y}(x,y)[\log(\bar{p}^{\rho}_{x|y}(x,y)) - \sum_s \bar{p}^{\rho}_{x|y}(s,y)\log(\bar{p}^{\rho}_{x|y}(s,y))][1+\log(\bar{p}^{\rho}_{x|y}(x,y))] \\
&= \frac{1}{1+\rho}\sum_x \bar{p}^{\rho}_{x|y}(x,y)\log(\bar{p}^{\rho}_{x|y}(x,y))[\log(\bar{p}^{\rho}_{x|y}(x,y)) - \sum_s \bar{p}^{\rho}_{x|y}(s,y)\log(\bar{p}^{\rho}_{x|y}(s,y))] \\
&= \frac{1}{1+\rho}\sum_x \bar{p}^{\rho}_{x|y}(x,y)\log(\bar{p}^{\rho}_{x|y}(x,y))\log(\bar{p}^{\rho}_{x|y}(x,y)) - \frac{1}{1+\rho}[\sum_x \bar{p}^{\rho}_{x|y}(x,y)\log(\bar{p}^{\rho}_{x|y}(x,y))]^2 \\
&\geq 0
\end{aligned} \tag{109}$$

The inequality is true by the Cauchy-Schwartz inequality and by noticing that $\sum_x \bar{p}^{\rho}_{x|y}(x,y) = 1$. ■

These properties will again be used in the proofs in the following lemmas.

Lemma 11: $\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial \rho} \geq 0$

Proof:

$$\begin{aligned}
\frac{\partial \frac{A(y,\rho)}{B(\rho)}}{\partial \rho} &= \frac{1}{B(\rho)^2}(\frac{\partial A(y,\rho)}{\partial \rho}B(\rho) - \frac{\partial B(\rho)}{\partial \rho}A(y,\rho)) \\
&= \frac{1}{B(\rho)^2}(A(y,\rho)H(\bar{p}^{\rho}_{x|y=y})B(\rho) - H(\bar{p}^{\rho}_{x|y})B(\rho)A(y,\rho)) \\
&= \frac{A(y,\rho)}{B(\rho)}(H(\bar{p}^{\rho}_{x|y=y}) - H(\bar{p}^{\rho}_{x|y}))
\end{aligned}$$

Now,

$$\begin{aligned}
\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial \rho} &= \frac{\partial}{\partial \rho}\sum_y \frac{A(y,\rho)}{B(\rho)}\sum_x \frac{C(x,y,\rho)}{D(y,\rho)}[-\log(\frac{C(x,y,\rho)}{D(y,\rho)})] \\
&= \frac{\partial}{\partial \rho}\sum_y \frac{A(y,\rho)}{B(\rho)}H(\bar{p}^{\rho}_{x|y=y}) \\
&= \sum_y \frac{A(y,\rho)}{B(\rho)}\frac{\partial H(\bar{p}^{\rho}_{x|y=y})}{\partial \rho} + \sum_y \frac{\partial \frac{A(y,\rho)}{B(\rho)}}{\partial \rho}H(\bar{p}^{\rho}_{x|y=y}) \\
&\geq \sum_y \frac{\partial \frac{A(y,\rho)}{B(\rho)}}{\partial \rho}H(\bar{p}^{\rho}_{x|y=y}) \\
&= \sum_y \frac{A(y,\rho)}{B(\rho)}(H(\bar{p}^{\rho}_{x|y=y}) - H(\bar{p}^{\rho}_{x|y}))H(\bar{p}^{\rho}_{x|y=y}) \\
&= \sum_y \frac{A(y,\rho)}{B(\rho)}H(\bar{p}^{\rho}_{x|y=y})^2 - H(\bar{p}^{\rho}_{x|y})^2 \\
&= (\sum_y \frac{A(y,\rho)}{B(\rho)}H(\bar{p}^{\rho}_{x|y=y})^2)(\sum_y \frac{A(y,\rho)}{B(\rho)}) - H(\bar{p}^{\rho}_{x|y})^2 \\
&\geq_{(a)} (\sum_y \frac{A(y,\rho)}{B(\rho)}H(\bar{p}^{\rho}_{x|y=y}))^2 - H(\bar{p}^{\rho}_{x|y})^2 \\
&= 0
\end{aligned} \tag{110}$$

where (a) is again true by the Cauchy-Schwartz inequality. ■

Lemma 12: $\frac{\partial D(\bar{p}^{\rho}_{xy}\|p_{xy})}{\partial \rho} = \rho\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial \rho}$

Proof: As shown in Lemma 15 and Lemma 17 respectively:

$$D(\bar{p}^{\rho}_{xy}\|p_{xy}) = \rho H(\bar{p}^{\rho}_{x|y}) - \log(\sum_y(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho})$$

$$H(\bar{p}^{\rho}_{x|y}) = \frac{\partial \log(\sum_y(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho})}{\partial \rho}$$

We have:

$$\begin{aligned}
\frac{\partial D(\bar{p}^{\rho}_{xy}\|p_{xy})}{\partial \rho} &= H(\bar{p}^{\rho}_{x|y}) + \rho\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial \rho} - \frac{\partial \log(\sum_y(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho})}{\partial \rho} \\
&= H(\bar{p}^{\rho}_{x|y}) + \rho\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial \rho} - H(\bar{p}^{\rho}_{x|y}) \\
&= \rho\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial \rho}
\end{aligned} \tag{111}$$

■

Lemma 13: $sign\frac{\partial[D(\bar{p}^{\rho}_{xy}\|p_{xy})-H(\bar{p}^{\rho}_{x|y})]}{\partial\rho} = sign(\rho-1)$.

Proof: Using the previous lemma, we get:

$$\frac{\partial D(\bar{p}^{\rho}_{xy}\|p_{xy}) - H(\bar{p}^{\rho}_{x|y})}{\partial\rho} = (\rho-1)\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial\rho}$$

Then by Lemma 11, we get the conclusion. ■

Lemma 14:

$$\rho H(p^{\rho}_{xy}) - (1+\rho)\log(\sum_y\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}) = D(p^{\rho}_{xy}\|p_{xy})$$

Proof: By noticing that $\log(p_{xy}(x,y)) = (1+\rho)[\log(p^{\rho}_{xy}(x,y)) + \log(\sum_{s,t}p_{xy}(s,t)^{\frac{1}{1+\rho}})]$. We have:

$$\begin{aligned}
D(p^{\rho}_{xy}\|p_{xy}) &= -H(p^{\rho}_{xy}) - \sum_{x,y}p^{\rho}_{xy}(x,y)\log(p_{xy}(x,y)) \\
&= -H(p^{\rho}_{xy}) - \sum_{x,y}p^{\rho}_{xy}(x,y)(1+\rho)[\log(p^{\rho}_{xy}(x,y)) + \log(\sum_{s,t}p_{xy}(s,t)^{\frac{1}{1+\rho}})] \\
&= -H(p^{\rho}_{xy}) + (1+\rho)H(p^{\rho}_{xy}) - (1+\rho)\sum_{x,y}p^{\rho}_{xy}(x,y)\log(\sum_{s,t}p_{xy}(s,t)^{\frac{1}{1+\rho}}) \\
&= \rho H(p^{\rho}_{xy}) - (1+\rho)\log(\sum_{s,t}p_{xy}(s,t)^{\frac{1}{1+\rho}}) \qquad (112)
\end{aligned}$$

■

Lemma 15:

$$\rho H(\bar{p}^{\rho}_{x|y}) - \log(\sum_y(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho}) = D(\bar{p}^{\rho}_{xy}\|p_{xy})$$

Proof:

$$\begin{aligned}
D(\bar{p}^{\rho}_{xy}\|p_{xy}) &= \sum_y\sum_x\frac{A(y,\rho)}{B(\rho)}\frac{C(x,y,\rho)}{D(y,\rho)}\log(\frac{\frac{A(y,\rho)}{B(\rho)}\frac{C(x,y,\rho)}{D(y,\rho)}}{p_{xy}(x,y)}) \\
&= \sum_y\sum_x\frac{A(y,\rho)}{B(\rho)}\frac{C(x,y,\rho)}{D(y,\rho)}[\log(\frac{A(y,\rho)}{B(\rho)}) + \log(\frac{C(x,y,\rho)}{D(y,\rho)}) - \log(p_{xy}(x,y))] \\
&= -\log(B(\rho)) - H(\bar{p}^{\rho}_{x|y}) + \sum_y\sum_x\frac{A(y,\rho)}{B(\rho)}\frac{C(x,y,\rho)}{D(y,\rho)}[\log(D(y,\rho)^{1+\rho}) - \log(C(x,y,\rho)^{1+\rho})] \\
&= -\log(B(\rho)) - H(\bar{p}^{\rho}_{x|y}) + (1+\rho)H(\bar{p}^{\rho}_{x|y}) \\
&= -\log(\sum_y(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho}) + \rho H(\bar{p}^{\rho}_{x|y})
\end{aligned}$$

■

Lemma 16:

$$H(p^{\rho}_{xy}) = \frac{\partial(1+\rho)\log(\sum_y\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})}{\partial\rho}$$

Proof:

$$\begin{aligned}
&\frac{\partial(1+\rho)\log(\sum_y\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})}{\partial\rho}\\
&= \log(\sum_t\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}) - \sum_y\sum_x \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_t\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}}\log(p_{xy}(x,y)^{\frac{1}{1+\rho}})\\
&= -\sum_y\sum_x \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_t\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}}\log(\frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_t\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}})\\
&= H(p^\rho_{xy}) \qquad (113)
\end{aligned}$$

■

Lemma 17:

$$H(\bar{p}^\rho_{x|y}) = \frac{\partial\log(\sum_y(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho})}{\partial\rho}$$

Proof: Notice that $B(\rho) = \sum_y(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho}$, and $\frac{\partial B(\rho)}{\partial\rho} = B(\rho)H(\bar{p}^\rho_{x|y})$ as shown in Lemma 10. It is clear that:

$$\begin{aligned}
\frac{\partial\log(\sum_y(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho})}{\partial\rho} &= \frac{\partial\log(B(\rho))}{\partial\rho}\\
&= \frac{1}{B(\rho)}\frac{\partial B(\rho)}{\partial\rho}\\
&= H(\bar{p}^\rho_{x|y}) \qquad (114)
\end{aligned}$$

■

REFERENCES

[1] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, pp. 471–480, July 1973.

[2] P. Koulgi, E. Tuncel, S. Regunathan, and K. Rose, "On zero-error coding of correlated sources," *IEEE Trans. Inform. Theory*, vol. 49, pp. 2856–2873, Nov. 2003.

[3] A. Sahai and T. Şimşek, "On the variable-delay reliability function of discrete memoryless channels with access to noisy feedback," in *IEEE Information Theory Workshop, San Antonio, Texas*, 2004.

[4] S. C. Draper and A. Sahai, "Noisy feedback improves communication reliability," in *Proc. Int. Symp. Inform. Theory*, 2006.

[5] S. C. Draper, "Universal incremental slepian-wolf coding," in *Proc. 42nd Allerton Conf. on Communication, Control and Computing*, Oct. 2004.

[6] C. Chang and A. Sahai, "The error exponent with delay for lossless source coding," *IEEE Information Theory Workshop*, March 2006.

[7] F. Jelinek, "Buffer overflow in variable length coding of fixed rate sources," *IEEE Trans. Inform. Theory*, vol. 14, pp. 490–501, May 1968.

[8] I. Csiszár and J. Körner, *Information Theory, Coding Theorems for Discrete Memoryless Systems*. Akadémiai Kiadó, 1981.

[9] R. G. Gallager, "Source coding with side information and universal coding," Mass. Instit. Tech., Tech. Rep. LIDS-P-937, 1976.

[10] G. Forney, "Convolutional codes iii. sequential decoding," *Information and Control*, vol. 25, no. 3, pp. 267–297, 1974.

[11] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2148–2177, Oct. 1998.

[12] T. M. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Inform. Theory*, vol. 21, pp. 226–228, Mar. 1975.

[13] A. Sahai and S. Mitter, "Source coding and channel requirements for unstable processes," *Submitted to IEEE Trans. Inform. Theory*, 2006.

[14] C. Chang and A. Sahai, "Upper bound on error exponents with delay for lossless source coding with side-information," *Proc. Int. Symp. Inform. Theory*, July 2006.

[15] A. Sahai, "Why block length and delay are not the same thing," *Submitted to IEEE Trans. Inform. Theory*, 2006.

[16] L. Weng, S. Pradhan, and A. Anastasopoulos, "Error exponent regions for gaussian broadcast and mulitple access channels," *submitted to Transactions on Information Theory*, 2005.

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.